SCORE-BASED DIFFUSION MODELS VIA STOCHASTIC DIFFERENTIAL EQUATIONS – A TECHNICAL TUTORIAL

WENPIN TANG AND HANYANG ZHAO

ABSTRACT. This is an expository article on the score-based diffusion models, with a particular focus on the formulation via stochastic differential equations (SDE). After a gentle introduction, we discuss the two pillars in the diffusion modeling – sampling and score matching, which encompass the SDE/ODE sampling, score matching efficiency, the consistency models, and reinforcement learning. Short proofs are given to illustrate the main idea of the stated results. The article is primarily a technical introduction to the field, and practitioners may also find some analysis useful in designing new models or algorithms.

> *What I cannot create, I do not understand. – Richard Feynman*

1. INTRODUCTION

Diffusion models describe a family of generative models that genuinely create the desired target distribution from noise. Inspired from energy-based modeling [60], [27, 65, 67] formalized the idea of diffusion models, which underpins the recent success in the text-to-image creators such as DALL·E 2 [55] and Stable Diffusion [58], and the text-to-video generator Sora [51]. Roughly speaking, diffusion models rely on a forward-backward procedure:

- *Forward deconstruction*: starting from the target distribution $X_0 \sim p_{\text{data}}(\cdot)$, the model gradually adds noise to transform the signal into noise $X_0 \to X_1 \to \cdots \to X_n \sim p_{\text{noise}}(\cdot)$.
- *Backward construction*: start with the noise $X_n \sim p_{\text{noise}}(\cdot)$, and reverse the forward process to recover the signal from noise $X_n \to X_{n-1} \to \cdots \to X_0 \sim p_{\text{data}}(\cdot)$.

The forward deconstruction is straightforward. What's the key in the diffusion model is the backward construction, and the underlying question is how to reverse the forward process. The answer hinges on two pillars: *time reversal* of (Markov) diffusion processes to set the form of the backward process (Section 2), and *score matching* to learn this process (Section 4).

In the original work [27, 61, 65], the forward/backward processes are specified by discrete-time Markov chains; [64, 67] unified the previous models through the lens of stochastic differential equations (SDE). In fact, there is no conceptual distinction between discrete

(A) Original Swiss Roll

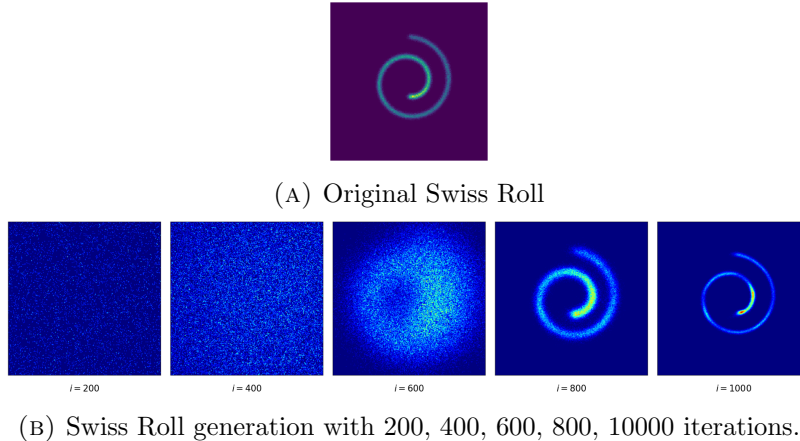(B) Swiss Roll generation with 200, 400, 600, 800, 10000 iterations.

FIGURE 1. Swiss Roll generation.

and continuous diffusion models because the diffusion models specified by the SDEs can be regarded as the continuum limits of the discrete models (Section 3), and the discrete diffusion models are obtained from the continuous models by suitable time discretization (Section 5.3). The view is that the SDEs unveil structural properties of the models, whereas the discrete counterparts are practical implementation.

The purpose of this article is to provide a tutorial on the recent theory of score-based diffusion models, mainly from a *continuous perspective* with a *statistical focus*. References on the discrete models will also be given. We sketch proofs for most stated results, and assumptions are given only when they are crucial in the analysis. We often use the phrase "under suitable conditions" to avoid less important technical details and keep the presentation concise and to the point. The paper serves as a gentle introduction to the field, and practitioners will find some analysis useful in designing new models or algorithms. Some results (e.g., in Sections 5.2, 6.2 and 7.3) appear for the first time here. Since the SDE formulation is adopted, we assume that the readers are familiar with basic stochastic calculus. Øksendal's book [50] provides a user-friendly account for stochastic analysis, and more advanced textbooks are [34, 68]. See also [76] for a literature review on diffusion models, and [8] for an optimization overview with more advanced materials such as diffusion guidance and fine-tuning.

The remainder of the paper is organized as follows. In Section 2, we start with the time reversal formula of diffusion processes, which is the cornerstone of diffusion models. Concrete examples are provided in Section 3. Section 4 is concerned with score matching techniques, another key ingredient of diffusion models. In Section 5, we consider the stochastic sampler of diffusion models, and analyze their convergence. In Section 6, a deterministic sampler – the probability ODE flow is introduced, along with its application to the consistency models. Additional results on score matching are given in Section 7. Concluding remarks and future directions are summarized in Section 8.

Notations: Below we collect a few notations that will be used throughout.

- For x, y vectors, denote by $x \cdot y$ the inner product between x and y, and $|x|$ the Euclidean (L^2) norm of x.

- For A a matrix (or a vector), $A^\top$ denotes the transpose of A. For A a square matrix, $\text{Tr}(A)$ is the trace of A.
- For f a vector field, ∇f is the gradient of f, $\nabla \cdot f = div(f)$ is the divergence of f, $\nabla^2 f$ is the Hessian of f, and Δf is the Laplacian of f. For $f = (f_1, \ldots, f_n)$ a matrix with f_i its ith column, $(\nabla f)_i^\top = \nabla f_i$ (the Jacobian matrix), and $(\nabla \cdot f)_i = div(f_i)$ for $i = 1, \ldots, n$.
- The symbol $a = \mathcal{O}(b)$ means that a/b is bounded, and $a = o(b)$ means that a/b tends to zero as some problem parameter tends to 0 or ∞.
- For X a random variable, $\mathcal{L}(X)$ denotes the probability distribution of X, and $X \sim p(\cdot)$ means that X has the distribution $p(\cdot)$. The notation $X \stackrel{d}{=} Y$ represents that X and Y have the same distribution.
- $\mathbb{E}X$ and $\text{Var}(X)$ denote the expectation and the variance of X respectively. We use $\widehat{\mathbb{E}}X$ for the empirical average of X, i.e. $\widehat{\mathbb{E}}X = \frac{1}{n}\sum_{i=1}^{n} X_i$ for $X_1, \cdots, X_n$ independent and identically distributed as X.
- $\mathcal{N}(\mu, \Sigma)$ denotes Gaussian distribution with mean μ and covariance matrix Σ.
- For P and Q two probability distributions, $d_{TV}(P, Q) = \sup_A |P(A) - Q(A)|$ is the total variation distance between P and Q; $\text{KL}(P, Q) = \int \log\left(\frac{dP}{dQ}\right) dP$ is the Kullback-Leilber (KL) divergence between P and Q; $W_2(P, Q) = \left(\inf_\gamma \mathbb{E}_{(X,Y)\sim\gamma}|X - Y|^2\right)^{\frac{1}{2}}$, where the infimum is taken over all couplings γ of P and Q, is the Wasserstein-2 distance between P and Q.

2. Time reversal formula

In this section, we present the time reversal formula, which lays the foundation for the diffusion models.

Consider a general (forward) SDE:

$$dX_t = f(t, X_t)dt + g(t, X_t)dB_t, \quad X_0 \sim p_{\text{data}}(\cdot), \tag{2.1}$$

where $(B_t, t \geq 0)$ is n-dimensional Brownian motion, and $f : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^d$ and $g : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^{d \times n}$ are diffusion parameters. Some conditions on $f(\cdot, \cdot)$ and $g(\cdot, \cdot)$ are required so that the SDE (2.1) is well-defined. For instance,

- If f and g are Lipschitz and have linear growth in x uniformly in t, then (2.1) has a strong solution which is pathwise unique.
- If f is bounded, and g is bounded, continuous and strictly elliptic, then (2.1) has a weak solution which is unique in distribution.

See [34, 50] for background on the well-posedness of SDEs, and [12, Chapter 1] for a review of related results.

For ease of presentation, we assume that X_t has a (suitably smooth) probability density $p(t, \cdot)$. Let $T > 0$ be fixed, and run the SDE (2.1) until time T to get $X_T \sim p(T, \cdot)$. Now if we start with $p(T, \cdot)$ and run the process X backward for time T, then we can generate copies of $p(0, \cdot) = p_{\text{data}}(\cdot)$. More precisely, consider the time reversal $Y_t := X_{T-t}$ for $0 \leq t \leq T$.

Assuming that Y also satisfies an SDE, we can run the backward process:

$$dY_t = \bar{f}(t, Y_t)dt + \bar{g}(t, Y_t)dB_t, \quad Y_0 \sim p(T, \cdot).$$

So we generate the desired $Y_T \sim p_{\text{data}}(\cdot)$ at time T.

As mentioned earlier, the high-level idea of diffusion models is to create the target distribution from noise. This means that the noise should not depend on the target distribution. Thus, we replace the (backward) initialization $Y_0 \sim p(T, \cdot)$ with some noise $p_{\text{noise}}(\cdot)$:

$$dY_t = \bar{f}(t, Y_t)dt + \bar{g}(t, Y_t)dB_t, \quad Y_0 \sim p_{\text{noise}}(\cdot), \tag{2.2}$$

as an approximation. Two natural questions arise:

(1) How do we choose the noise $p_{\text{noise}}(\cdot)$?
(2) What are the parameters $\bar{f}(\cdot, \cdot)$ and $\bar{g}(\cdot, \cdot)$?

For (1), the noise $p_{\text{noise}}(\cdot)$ is commonly derived by decoupling X_0 from the conditional distribution of $(X_T \mid X_0)$, as we will explain with the examples in Section 3. It is expected that the closer the the distributions $p(T, \cdot)$ and $p_{\text{noise}}(\cdot)$ are, the closer the distribution of Y_T sampled from (2.2) is to $p_{\text{data}}(\cdot)$. For (2), it relies on the following result on the time reversal of SDEs. To simplify the notations, we write

$$a(t, x) := g(t, x)g(t, x)^\top.$$

Theorem 2.1 (Time reversal formula). [1, 26] *Under suitable conditions on* $f(\cdot, \cdot)$, $g(\cdot, \cdot)$ *and* $\{p(t, \cdot)\}_{0 \le t \le T}$, *we have:*

$$\bar{g}(t, x) = g(T - t, x), \quad \overline{f}(t, x) = -f(T - t, x) + \frac{\nabla \cdot (p(T - t, x)a(T - t, x))}{p(T - t, x)}. \tag{2.3}$$

Proof. We give a heuristic derivation of the formula (2.3). First, the infinitesimal generator of X is $\mathcal{L} := \frac{1}{2}\nabla \cdot a(t, x)\nabla + f_a \cdot \nabla$, where $f_a := f - \frac{1}{2}\nabla \cdot a$. It is known that the density $p(t, x)$ satisfies the Fokker–Planck equation:

$$\frac{\partial}{\partial t}p(t, x) = \mathcal{L}^*p(t, x), \tag{2.4}$$

where $\mathcal{L}^* := \frac{1}{2}\nabla \cdot a(t, x)\nabla - \nabla \cdot f_a$ is the adjoint operator of $\mathcal{L}$. Let $\bar{p}(t, x) := p(T - t, x)$ be the probability density of the time reversal Y. By (2.4), we get

$$\frac{\partial}{\partial t}\bar{p}(t, x) = -\frac{1}{2}\nabla \cdot (a(T - t, x)\nabla\bar{p}(t, x)) + \nabla \cdot (f_a(T - t, x)\bar{p}(t, x)). \tag{2.5}$$

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

$$\frac{\partial}{\partial t}\bar{p}(t, x) = \frac{1}{2}\nabla \cdot (\bar{a}(t, x)\nabla\bar{p}(t, x)) - \nabla \cdot (\overline{f_{\bar{a}}}(t, x)\bar{p}(t, x)). \tag{2.6}$$

Comparing (2.5) and (2.6), we set $\bar{a}(t, x) = a(T - t, x)$ and then get

$$(f_a(T - t, x) + \overline{f_{\bar{a}}}(t, x))\bar{p}(t, x) = a(T - t, x)\nabla\bar{p}(t, x),$$

which can be rewritten as:

$$(f(T - t, x) + \overline{f}(t, x))\bar{p}(t, x) - \nabla \cdot a(T - t, x)\bar{p}(t, x) = a(T - t, x)\nabla\bar{p}(t, x).$$

This yields the desired result. $\qquad\square$

Let's comment on Theorem 2.1. [26, 49] proved the result by assuming that $f(\cdot, \cdot)$ and $g(\cdot, \cdot)$ 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 [54], where only the boundedness of $\nabla a(t, x)$ in some L^2 norm is required. In another direction, [20, 21] used an entropy argument to prove the time reversal formula in the case of $n = d$ and $g(t, x) = \sigma I$. This approach was further developed by [6] in connection with optimal transport theory.

By Theorem 2.1, the backward process is:

$$dY_t = (-f(T - t, Y_t) + a(T - t, Y_t)\nabla \log p(T - t, Y_t) + \nabla \cdot a(T - t, Y_t))dt$$
$$+ g(T - t, Y_t)dB_t, \quad Y_0 \sim p_{\text{noise}}(\cdot). \tag{2.7}$$

Since $f(\cdot, \cdot)$ and $g(\cdot, \cdot)$ are chosen in advance, all but the term $\nabla \log p(T - t, Y_t)$ in (2.7) are avaliable. So in order to run the backward process (2.7), we need to learn $\nabla \log p(t, x)$, known as *Stein's score function*. Recently developed score-based generative modeling consists of estimating $\nabla \log p(t, x)$ by function approximations, which will be discussed in Section 4.

3. Examples

We have seen that a diffusion model is specified by the pair $(f(\cdot, \cdot), g(\cdot, \cdot))$. The design of $(f(\cdot, \cdot), g(\cdot, \cdot))$ is important because it determines the quality of data generation. There are two general rules of thumb – easy learning from the forward process, and good sampling from the backward process, which will be clear in the next two sections.

Now let's provide some examples of the diffusion model (2.1)–(2.7). Most existing models take the form:

$$n = d \quad \text{and} \quad g(t, x) = g(t)I,$$

where $g(t) \in \mathbb{R}_+$. That is, the model parameter $g(\cdot, \cdot)$ is only time-dependent, rather than (time and) state-dependent. One important reason is that for the SDEs with state-dependent coefficient, it is often not easy to decouple X_0 from the distribution of $(X_T \,|\, X_0)$. Hence, it is not clear how to pick the noise $p_{\text{noise}}(\cdot)$ as a proxy to $X_T \sim p(T, \cdot)$, as is the case of geometric Brownian motion.

(a) *Ornstein-Ulenback (OU) process* [14]:

$$f(t, x) = \theta(\mu - x) \text{ with } \theta > 0, \, \mu \in \mathbb{R}^d \quad \text{and} \quad g(t) = \sigma > 0. \tag{3.1}$$

Given $X_0 = x$, the distribution of X_t is $p(t, \cdot; x) = \mathcal{N}(\mu + (x - \mu)e^{-\theta t}, \frac{\sigma^2}{2\theta}(1 - e^{-2\theta t})I)$. An obvious candidate for the noise is the stationary distribution of the OU process $p_{\text{noise}}(\cdot) = \mathcal{N}(\mu, \frac{\sigma^2}{2\theta}I)$. The backward process is specified to:

$$dY_t = \left(\theta(Y_t - \mu) + \sigma^2\nabla \log p(T - t, Y_t)\right)dt + g\,dB_t, \quad Y_0 \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{2\theta}I\right). \tag{3.2}$$

More generally, consider the overdamped Langevin process:

$$f(t, x) = -\nabla U(x) \quad \text{and} \quad g(t) = \sigma > 0, \tag{3.3}$$

with a suitable landscape $U : \mathbb{R}^d \to \mathbb{R}$. The OU process corresponds to the choice of $U(x) = \frac{\theta}{2}|x - \mu|^2$. The backward process is then:

$$dY_t = \left(\nabla U(Y_t) + \sigma^2 \nabla \log p(T - t, Y_t)\right) dt + \sigma dB_t, \quad Y_0 \propto \exp\left(-\frac{2\,U(\cdot)}{\sigma^2}\right). \tag{3.4}$$

(b) *Variance exploding (VE) SDE* [67]: this is the continuum limit of *score matching with Langevin dynamics* (SMLD) [65]. The idea of SMLD is to use N noise scales $\sigma_0 < \sigma_1 < \cdots < \sigma_{N-1}$, and run the (forward) Markov chain:

$$x_i = x_{i-1} + \sqrt{\sigma_i^2 - \sigma_{i-1}^2}\, z_{i-1}, \quad 1 \le i \le N,$$

where $z_0, \ldots, z_{N-1}$ are independent and identically distributed $\mathcal{N}(0, I)$. By taking $\Delta t = \frac{T}{N} \to 0$, $\sigma(i\Delta t) = \sigma_i$, $x_{i\Delta t} = x_i$ and $z_{i\Delta t} = z_i$, we get $x_{t+\Delta t} = x_t + \sqrt{\sigma^2(t + \Delta t) - \sigma^2(t)}\, z_t \approx x_t + \sqrt{\frac{d\sigma^2(t)}{dt}\Delta t}\, z_t$. That is,

$$dX_t = \sqrt{\frac{d\sigma^2(t)}{dt}}\, dB_t, \quad 0 \le t \le T.$$

This implies that $f(t, x) = 0$ and $g(t) = \sqrt{\frac{d\sigma^2(t)}{dt}}$. The noise scales are typically set to be a geometric sequence $\sigma(t) = \sigma_{\min}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{\frac{t}{T}}$, with $\sigma_{\min} \ll \sigma_{\max}$. Thus,

$$f(t, x) = 0 \quad \text{and} \quad g(t) = \sigma_{\min}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{\frac{t}{T}} \sqrt{\frac{2}{T} \log \frac{\sigma_{\max}}{\sigma_{\min}}}. \tag{3.5}$$

Here $X_t = X_0 + \int_0^t g(s)dB_s$ is the Paley-Wiener integral. Given $X_0 = x$, the distribution of X_t is:

$$p(t, \cdot; x) = \mathcal{N}\left(x, \left(\int_0^t g^2(s)ds\right) I\right) = \mathcal{N}\left(x, \sigma_{\min}^2\left(\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{\frac{2t}{T}} - 1\right) I\right). \tag{3.6}$$

The name "variance exploding" comes from the fact that $\text{Var}(X_0) \ll \text{Var}(X_T)$ because $\sigma_{\min} \ll \sigma_{\max}$. Note that the forward process X does not have a stationary distribution, but we can decouple x from $p(T, \cdot) = \mathcal{N}\left(x, (\sigma_{\max}^2 - \sigma_{\min}^2)I\right)$ to get

$$p_{\text{noise}}(\cdot) = \mathcal{N}(0, (\sigma_{\max}^2 - \sigma_{\min}^2)I)).$$

The backward process is:

$$dY_t = g^2(T - t))\nabla \log p(T - t, Y_t) + g(T - t)dB_t, \quad Y_0 \sim \mathcal{N}(0, (\sigma_{\max}^2 - \sigma_{\min}^2)I)), \tag{3.7}$$

where $g(t)$ is defined by (3.5).

As we will see in Section 6.2, the recently proposed consistency models [63] use $\sigma(t) = t$ and $g(t) = \sqrt{2t}$ suggested by [35], with $p_{\text{noise}}(\cdot) = \mathcal{N}(0, T^2 I)$ as an alternative parametrization.

(c) *Variance preserving (VP) SDE* [67]: this is the continuum limit of *denoising diffusion probabilistic models* (DDPM) [27]. DDPM uses N noise scales $\beta_1 < \beta_2 < \cdots < \beta_N$, and runs the (forward) Markov chain:

$$x_i = \sqrt{1 - \beta_i}x_{i-1} + \sqrt{\beta_i}z_i, \quad 1 \le i \le N.$$

Similarly, by taking the limit $\Delta t = \frac{T}{N} \to 0$ with $\beta(i\Delta t) = N\beta_i/T$, $x_{i\Delta t} = x_i$ and $z_{i\Delta t} = z_i$, we get $x_{t+\Delta t} = \sqrt{1 - \beta(t)\Delta t}\, x_t + \sqrt{\beta(t)\Delta t}\, z_t \approx x_t - \frac{1}{2}\beta(t)x_t\Delta t + \sqrt{\beta(t)\Delta t}\, z_t$. This leads to:

$$dX_t = -\frac{1}{2}\beta(t)X_t\, dt + \sqrt{\beta(t)}\, dB_t, \quad 0 \le t \le T.$$

We have:

$$f(t,x) = -\frac{1}{2}\beta(t)x \quad \text{and} \quad g(t) = \sqrt{\beta(t)}. \tag{3.8}$$

The noise scales of DDPM are typically an arithmetic sequence:

$$\beta(t) = \beta_{\min} + \frac{t}{T}(\beta_{\max} - \beta_{\min}), \quad \text{with } \beta_{\min} \ll \beta_{\max}, \tag{3.9}$$

(and $\beta_{\min}$, $\beta_{\max}$ are scaled to the order N/T.) By applying Itô's formula to $e^{\frac{1}{2}\int_0^t \beta(s)ds}X_t$, we get the distribution of $(X_t \,|\, X_0 = x)$:

$$p(t,\cdot\,;x) = \mathcal{N}\left(e^{-\frac{1}{2}\int_0^t \beta(s)ds}x, (1 - e^{-\int_0^t \beta(s)ds})I\right). \tag{3.10}$$

Thus, $p(T,\cdot\,;x) = \mathcal{N}(e^{-\frac{T}{4}(\beta_{\max}+\beta_{\min})}x, (1 - e^{-\frac{T}{2}(\beta_{\max}+\beta_{\min})})I)$, which is close to $\mathcal{N}(0,I)$ if $\beta_{\max}$, or T is set to be large. This justifies the name "variance preserving", and we can set $p_{\text{noise}} = \mathcal{N}(0,I)$. The backward process is:

$$dY_t = \left(\frac{1}{2}\beta(T-t)Y_t + \beta(T-t)\nabla \log p(T-t, Y_t)\right)dt + \sqrt{\beta(T-t)}dB_t, \quad Y_0 \sim \mathcal{N}(0,I), \tag{3.11}$$

where $\beta(t)$ is defined by (3.9).

Note that the choice of $\beta(t) = \sigma^2$ leads to the OU process with $f(t,x) = -\frac{1}{2}\sigma^2 x$ and $g(t) = \sigma$. In fact, the VP SDE is also referred to as the OU process in the literature [10, 40, 41]. Here we emphasize the difference between the OU process, and the VP SDE with the (typical) choice (3.9) for $\beta(t)$.

(d) *Sub-variance preserving (subVP) SDE* [67]:

$$f(t,x) = -\frac{1}{2}\beta(t)x \quad \text{and} \quad g(t) = \sqrt{\beta(t)(1 - e^{-2\int_0^t \beta(s)ds})}, \tag{3.12}$$

where $\beta(t)$ is defined by (3.9). The same reasoning as in (c) shows that given $X_0 = x$, the distribution of X_t is:

$$\begin{aligned}
p(t,\cdot\,;x) &= \mathcal{N}\left(e^{-\frac{1}{2}\int_0^t \beta(s)ds}x, (1 - e^{-\int_0^t \beta(s)ds})^2 I\right) \\
&= \mathcal{N}\left(e^{-\frac{t^2}{4T}(\beta_{\max}-\beta_{\min})-\frac{t}{2}\beta_{\min}}x, (1 - e^{-\frac{t^2}{2T}(\beta_{\max}-\beta_{\min})-t\beta_{\min}})^2 I\right).
\end{aligned} \tag{3.13}$$

Note that $\text{Var}_{\text{sub-VP}}(X_t) \le \text{Var}_{\text{VP}}(X_t)$, hence the name "sub-VP", and $p_{\text{noise}}(\cdot) = \mathcal{N}(0,I)$. Set $\gamma(t) := e^{-2\int_0^t \beta(s)ds} = e^{-\frac{t^2}{T}(\beta_{\max}-\beta_{\min})-2t\beta_{\min}}$, so $g(t) = \sqrt{\beta(t)(1-\gamma(t))}$. The backward process is:

$$\begin{aligned}
dY_t = &\left(\frac{1}{2}\beta(T-t)Y_t + \beta(T-t)(1 - \gamma(T-t))\nabla \log p(T-t, Y_t)\right)dt \\
&+ \sqrt{\beta(T-t)(1 - \gamma(T-t))}dB_t, \quad Y_0 \sim \mathcal{N}(0,I).
\end{aligned} \tag{3.14}$$

(e) *Contractive diffusion probabilistic models* (CDPM) [70]: the idea of CDPM is to force contraction on the backward process to narrow the score matching errors (at the cost of possible noise approximation bias). A practical criterion is:

$$(x - x') \cdot (f(t, x) - f(t, x')) \geq r_f(t)|x - x|^2, \quad \text{with} \inf_{t \in [0,T]} r_f(t) > 0. \tag{3.15}$$

Two examples are of interest: contractive OU process and contractive sub-VP SDE.

(e1) *Contractive Ornstein-Ulenback (COU) process*:

$$f(t, x) = \theta(x - \mu) \text{ with } \theta > 0, \nu \in \mathbb{R}^d \quad \text{and} \quad g(t) = \sigma > 0. \tag{3.16}$$

Similar to (a), we set $p_{\text{noise}}(\cdot) = \mathcal{N}\left(0, \frac{\sigma^2}{2\theta}(e^{2\theta T} - 1)I\right)$ by decoupling x from the conditional distribution of $(X_t \,|\, X_0 = x)$. The backward process is:

$$dY_t = \left(-\theta(Y_t - \mu) + \sigma^2 \nabla \log p(T - t, Y_t)\right) dt + \sigma dB_t, \quad Y_0 \sim \mathcal{N}\left(0, \frac{\sigma^2}{2\theta}(e^{2\theta T} - 1)I\right). \tag{3.17}$$

(e2) *Contractive variance preserving (CVP) SDE*:

$$f(t, x) = \frac{1}{2}\beta(t)x \quad \text{and} \quad g(t) = \sqrt{\beta(t)}, \tag{3.18}$$

where $\beta(t)$ is defined by (3.9). Similar to (c), we set $p_{\text{noise}}(\cdot) = \mathcal{N}\left(0, (e^{\frac{T}{2}(\beta_{\max} + \beta_{\min})} - 1)I\right)$. The backward process is:

$$dY_t = \left(-\frac{1}{2}\beta(T - t)Y_t + \beta(T - t)\nabla \log p(T - t, Y_t)\right) dt$$
$$+ \sqrt{\beta(T - t)}dB_t, \quad Y_0 \sim \mathcal{N}\left(0, (e^{\frac{T}{2}(\beta_{\max} + \beta_{\min})} - 1)I\right). \tag{3.19}$$

The hyperparameters of CDPM often needs to be fine-tuned with respect to T to trade off the noise approximation bias and the score matching errors.

4. Score matching techniques

In Section 2, we have seen that the main obstacle to implement the diffusion model (2.1)–(2.7) is the unknown score function $\nabla \log p(t, x)$. This section reviews the recently developed score-based generative modeling [27, 65, 67], whose goal is to estimate the score function $\nabla \log p(t, x)$ by a family of functions $\{s_\theta(t, x)\}_\theta$ (e.g. kernels and neural nets). This technique is referred to as *score matching*. With the (true) score function $\nabla \log p(t, x)$ being replaced with the score matching function $s_\theta(t, x)$, the backward process (2.7) becomes:

$$dY_t = (-f(T - t, Y_t) + a(T - t, Y_t)s_\theta(T - t, Y_t) + \nabla \cdot a(T - t, Y_t))dt$$
$$+ g(T - t, Y_t)dB_t, \quad Y_0 \sim p_{\text{noise}}(\cdot). \tag{4.1}$$

In the case of $g(t, x) = g(t)I$, it simplifies to:

$$dY_t = \left(-f(T - t, Y_t) + g^2(T - t)s_\theta(T - t, Y_t)\right) dt + g(T - t)dB_t, \quad Y_0 \sim p_{\text{noise}}(\cdot). \tag{4.2}$$

The plan of this section is as follows. In Section 4.1, we present the general score matching technique. The two (popularly used) scalable score matching methods – *sliced score matching* and *denoising score matching* will be discussed in Sections 4.2 and 4.3.

4.1. **Score matching.** Recall that $\{s_\theta(t,x)\}_\theta$ is a family of functions on $\mathbb{R}_+ \times \mathbb{R}^d$ parametrized by θ, which are used to approximate the score function $\nabla \log p(t,x)$. Fix time t, the goal is to solve the stochastic optimization problem:

$$\min_\theta \mathcal{J}_{\text{ESM}}(\theta) := \mathbb{E}_{p(t,\cdot)}|s_\theta(t,X) - \nabla \log p(t,X)|^2, \tag{4.3}$$

where $\mathbb{E}_{p(t,\cdot)}$ denotes the expectation taken over $X \sim p(t,\cdot)$. But the problem (4.3), known as the *explicit score matching* (ESM), is far-fetched because the score $\nabla \log p(t,X)$ on the right side is not available.

Interestingly, this problem has been studied in the context of estimating statistical models with unknown normalizing constant. (In fact, if $p(\cdot)$ is a Gibbs measure, then its score $\nabla \log p(\cdot)$ does not depend on the normalizing constant.) The following result shows that the score matching problem (4.3) can be recast into a feasible stochastic optimization with no $\nabla \log p(t,X)$-term, referred to as the *implicit score matching* (ISM).

Theorem 4.1 (Implicit score matching). [30] *Let*

$$\mathcal{J}_{ISM}(\theta) := \mathbb{E}_{p(t,\cdot)} \left[|s_\theta(t,X)|^2 + 2\nabla \cdot s_\theta(t,X) \right]. \tag{4.4}$$

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

Proof. We have

$$\nabla_\theta \mathcal{J}_{\text{ISM}}(\theta) = \nabla_\theta \mathbb{E}_{p(t,\cdot)} \left[|s_\theta(t,X)|^2 \right] - 2\mathbb{E}_{p(t,\cdot)} \left[\nabla_\theta s_\theta(t,X)^\top \nabla \log p(t,X) \right]$$

$$= \nabla_\theta \mathbb{E}_{p(t,\cdot)} \left[|s_\theta(t,X)|^2 \right] - 2\int \nabla_\theta s_\theta(t,x)^\top \nabla p(t,x) dx$$

$$= \nabla_\theta \mathbb{E}_{p(t,\cdot)} \left[|s_\theta(t,X)|^2 \right] - 2\nabla_\theta \int s_\theta(t,x)^\top \nabla p(t,x) dx$$

$$= \nabla_\theta \mathbb{E}_{p(t,\cdot)} \left[|s_\theta(t,X)|^2 \right] + 2\nabla_\theta \int \nabla \cdot s_\theta(t,x)\, p(t,x) dx$$

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

where we use the divergence theorem in the fourth equation. $\qquad \square$

Clearly, the implicit score matching problem (4.4) can be solved by stochastic optimization tools, e.g. stochastic gradient descent (SGD). Here the distribution $p(t,\cdot)$ is sampled by first picking a random data point from $p_{\text{data}}(\cdot)$, followed by the conditional distribution of $(X_t \,|\, X_0)$. This requires the distribution of $(X_t \,|\, X_0)$ be easy to sample – the easy learning criterion for diffusion models. Notably, the distribution of $(X_t \,|\, X_0)$ is Gaussian for all the examples in Section 3.

In practice, a time-weighted version of the problem (4.3) is considered:

$$\min_\theta \tilde{\mathcal{J}}_{\text{ESM}}(\theta) := \mathbb{E}_{t \in \mathcal{U}(0,T)} \mathbb{E}_{p(t,\cdot)} \left[\lambda(t)|s_\theta(t,X) - \nabla \log p(t,X)|^2 \right]$$

$$= \frac{1}{T} \int_0^T \mathbb{E}_{p(t,\cdot)} \left[\lambda(t)|s_\theta(t,X) - \nabla \log p(t,X)|^2 \right] dt, \tag{4.5}$$

where $\mathcal{U}(0,T)$ denotes the uniform distribution over $[0,T]$, and $\lambda : \mathbb{R} \to \mathbb{R}_+$ is a weight function. The corresponding implicit score matching problem is:

$$\min_\theta \tilde{\mathcal{J}}_{\mathrm{ISM}}(\theta) = \mathbb{E}_{t \in \mathcal{U}(0,T)} \mathbb{E}_{p(t,\cdot)} \left[\lambda(t)(|s_\theta(t,X)|^2 + 2\,\nabla \cdot s_\theta(t,X)) \right]. \tag{4.6}$$

Note, however, that the problem (4.4) or (4.6) can still be computationally costly when the dimension d is large. For instance, if we use a neural net as the function class of $s_\theta(t,x)$, we need to perform d times backpropagation of all the parameters to compute $\nabla \cdot s_\theta(t,x) = \mathrm{Tr}(\nabla s_\theta(t,x))$. This means that the computation of the derivatives scales linearly with the dimension, hence making the gradient descent methods inefficient to solve the score matching problem with respect to high dimensional data. Two alternative approaches are commonly used to deal with the scalability issue, which will be the focus of the next subsections.

4.2. **Sliced score matching.** The burden of computation comes from the term $\nabla \cdot s_\theta(t,x)$. One clever idea proposed by [66] is to tackle this term by random projections. It relies on the key observation:

$$\nabla \cdot s_\theta(t,x) = \mathbb{E}_{v \sim \mathcal{N}(0,I)} \left[v^\top \nabla s_\theta(t,x) v \right], \tag{4.7}$$

where $\nabla s_\theta(t,x) \in \mathbb{R}^{d \times d}$ is the Jacobian matrix of $s_\theta(t,x)$. The (implicit) score matching problem (4.6) can then be rewritten as:

$$\min_\theta \tilde{\mathcal{J}}_{\mathrm{SSM}}(\theta) = \mathbb{E}_{t \in \mathcal{U}(0,T)} \mathbb{E}_{v \sim \mathcal{N}(0,I)} \mathbb{E}_{p(t,\cdot)} \left[\lambda(t) \left(|s_\theta(t,X)|^2 + 2\,v^\top \nabla(v^\top s_\theta(t,x)) \right) \right], \tag{4.8}$$

which is referred to as the *sliced score matching* (SSM). Since $\tilde{\mathcal{J}}_{\mathrm{SSM}}(\theta) = \tilde{\mathcal{J}}_{\mathrm{ISM}}(\theta)$ for all θ, we have the following result.

Theorem 4.2 (Sliced score matching). [66] *Under suitable conditions on s_θ, we have $\tilde{\mathcal{J}}_{SSM}(\theta) = \tilde{\mathcal{J}}_{ESM}(\theta) + C$ for some C independent of θ. Consequently, the minimum point of $\tilde{\mathcal{J}}_{SSM}(\theta)$ and that of $\tilde{\mathcal{J}}_{ESM}(\theta)$ coincide.*

Note that for a single fixed v, it only requires one-time backpropagation because the term $v^\top s_\theta(t,x)$ can be regarded as adding a layer of the inner product between v and $s_\theta(t,x)$. To get the expectation $\mathbb{E}_{v \sim \mathcal{N}(0,I)}$ in (4.8), we pick m samples of $v_i \sim \mathcal{N}(0,I)$, $1 \leq i \leq m$, compute the objective for each v_i, and then take the average. So it requires m times backpropagation. Typically, m is set to be small ($m \ll d$), and empirical studies show that $m = 1$ is often good enough.

4.3. **Denoising score matching.** The second approach relies on conditioning X_t on $X_0 \sim p_{data}(\cdot)$, known as denoising score matching (DSM) [30, 73]. Let's go back to the ESM problem (4.5), and it is equivalent to the following DSM problem:

$$\tilde{\mathcal{J}}_{\mathrm{DSM}}(\theta) = \mathbb{E}_{t \sim \mathcal{U}(0,T)} \left\{ \lambda(t) \mathbb{E}_{X_0 \sim p_{data}(\cdot)} \mathbb{E}_{X_t | X_0} \left[|s_\theta(t,X_t) - \nabla \log p(t, X_t | X_0)|^2 \right] \right\}, \tag{4.9}$$

where the gradient $\nabla \log p(t, X_t | X_0)$ is with respect to X_t.

Theorem 4.3 (Denoising score matching). [73] *Under suitable conditions on s_θ, we have $\tilde{\mathcal{J}}_{DSM}(\theta) = \tilde{\mathcal{J}}_{ESM}(\theta) + C$ for some C independent of θ. Consequently, the minimum point of $\tilde{\mathcal{J}}_{DSM}$ and that of $\tilde{\mathcal{J}}_{ESM}$ coincide.*

Proof. Let $\mathcal{J}_{\mathrm{DSM}}(\theta) := \mathbb{E}_{X_0 \sim p_{data}(\cdot)} \mathbb{E}_{X_t|X_0} \left[|s_\theta(t, X_t) - \nabla \log p(t, X_t \mid X_0)|^2 \right]$. It suffices to prove that $\mathcal{J}_{\mathrm{DSM}}(\theta) = \mathcal{J}_{\mathrm{ESM}}(\theta) + C$ for some C independent of θ. Note that

$$\mathcal{J}_{\mathrm{ESM}}(\theta) = \mathbb{E}_{p(t,\cdot)} |s_\theta(t, X) - \nabla \log p(t, X)|^2$$
$$= \mathbb{E}_{p(t,\cdot)} \left[|s_\theta(t, X)|^2 - 2 s_\theta(t, X)^\top \nabla \log p(t, X) + |\nabla \log p(t, X)|^2 \right].$$

For the inner product, we can rewrite it as:

$$\mathbb{E}_{p(t,\cdot)} \left[s_\theta(t, X)^\top \nabla \log p(t, X) \right] = \int_x s_\theta(t, x)^\top \nabla p(t, x) dx$$
$$= \int_x s_\theta(t, x)^\top \nabla \int_{x_0} p(0, x_0) p(t, x; x_0) dx_0 dx$$
$$= \int_{x_0} \int_x s_\theta(t, x)^\top p(0, x_0) \nabla p(t, x; x_0) dx dx_0$$
$$= \int_{x_0} p(0, x_0) \int_x s_\theta(t, x)^\top p(t, x; x_0) \nabla \log p(t, ; | x_0) dx dx_0$$
$$= \mathbb{E}_{X_0 \sim p_{data}(\cdot)} \mathbb{E}_{X_t|X_0} \left[s_\theta(t, X_t)^\top \nabla \log p(t, X_t \mid X_0) \right].$$

Combining with $\mathbb{E}_{p(t,\cdot)} |s_\theta(t, X)|^2 = \mathbb{E}_{X_0 \sim p_{data}(\cdot)} \mathbb{E}_{X_t|X_0} |s_\theta(t, X)|^2$ concludes the proof. $\square$

The main takeaway of DSM is that the gradient of the log density at some corrupted point should ideally move towards the clean sample. As mentioned earlier, the conditional distribution of $(X_t \mid X_0)$ is required to be simple, e.g. Gaussian. Here we set $(X_t \mid X_0) \sim \mathcal{N}(\mu_t(X_0), \sigma_t^2 I)$ for some $\mu_t(\cdot)$ and $\sigma_t > 0$. For instance, $\mu_t(x) = x$ and $\sigma_t = \sigma_{\min} \sqrt{\left(\frac{\sigma_{\max}}{\sigma_{\min}} \right)^{\frac{2t}{T}} - 1}$ for VE, and $\mu(t, x) = e^{-\frac{1}{2} \int_0^t \beta(s) ds} x$ and $\sigma_t = e^{-\frac{1}{2} \int_0^t \beta(s) ds}$ for VP. In this case, we can compute explicitly the conditional score:

$$\nabla \log p(t, X_t \mid X_0) = \frac{\mu_t(X_0) - X_t}{\sigma_t^2}. \tag{4.10}$$

The direction $\frac{1}{\sigma_t^2} (X_t - \mu_t(X_0))$ clearly facilitates moving to the clean sample, and we want $s_\theta(t, x)$ to match the score (4.10) as best it can. Empirically, it was observed [67] that a good candidate for the weight function $\lambda(t)$ is:

$$\lambda(t) \propto 1/\mathbb{E} |\nabla \log p(t, X_t \mid X_0)|^2 = \sigma_t^2. \tag{4.11}$$

The choice (4.11) is related to an evidence lower bound (ELBO) of the KL divergence between the generated distribution and the true distribution, see [29, 47, 64] for discussions. Injecting (4.11) into (4.9) yields the ultimate objective:

$$\tilde{\mathcal{J}}_{\mathrm{DSM}}(\theta) = \mathbb{E}_{t \sim \mathcal{U}(0,T)} \left\{ \sigma_t^2 \mathbb{E}_{X_0 \sim p_{data}(\cdot)} \mathbb{E}_{X_t|X_0} \left[\left| s_\theta(t, X_t) + \frac{X_t - \mu_t(X_0)}{\sigma_t^2} \right|^2 \right] \right\}$$
$$= \mathbb{E}_{t \sim \mathcal{U}(0,T)} \left\{ \mathbb{E}_{X_0 \sim p_{data}(\cdot)} \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)} \left[|\sigma_t s_\theta(t, \mu_t(X_0) + \sigma_t \epsilon) + \epsilon|^2 \right] \right\}, \tag{4.12}$$

where the second equality follows from a change of variables $\epsilon := \frac{X_t - \mu_t(X_0)}{\sigma_t}$.

5. Stochastic samplers: convergence results

In this section, we consider the convergence of the stochastic samplers, i.e., the backward processes, under a "blackbox" assumption on score matching:

Assumption 5.1. *There exists $\varepsilon > 0$ such that $\mathbb{E}_{p(t,\cdot)}|s_\theta(t, X) - \nabla \log p(t, X)|^2 \leq \varepsilon^2$, $t \in [0, T]$ for some $s_\theta(\cdot, \cdot)$.*

For instance, the score matching function $s_\theta(\cdot, \cdot)$ can be the output of any algorithm introduced in Section 4. This condition postulates an L^2-bound on score matching, which is implied by the stronger L^∞-bound:

Assumption 5.1'. *There exists $\varepsilon > 0$ such that $|s_\theta(t, \cdot) - \nabla \log p(t, \cdot)|_\infty \leq \varepsilon$, $t \in [0, T]$ for some $s_\theta(\cdot, \cdot)$.*

As we will see in Section 7, there has been a body of work on the accuracy of score matching. So these assumptions, especially Assumption 5.1 can be replaced with those score approximation bounds.

The quality of the diffusion model relies on how $(f(\cdot, \cdot), g(\cdot))$ is chosen. Since the SDEs are inherently in continuous time, the convergence of the diffusion model as a continuous process yields another criterion on the design of the model. Ideally, we want to show that the stochastic sampler Y_T defined by (4.1), as well as its discretization are close in distribution to the target distribution $p_{\text{data}}(\cdot)$. Here we focus on two metrics: the total variation distance and the Wasserstein-2 distance. While most existing convergence results were established in the total variation distance (or KL divergence), the convergence in the Wasserstein metric is important because it has shown to align with human judgment on image similarity [5], and the standard evaluation metric – Fréchet inception distance (FID) is also based on the Wasserstein distance.

For simplicity, we assume (as in Section 3) that $n = d$ and $g(t, x) = g(t)I$. In Sections 5.1–5.2, we study the convergence of the stochastic samplers in continuous time in the two metrics respectively. The problem of discretization will be discussed in Section 5.3.

5.1. **Total variation bound.** The goal is to bound $d_{TV}(\mathcal{L}(Y_T), p_{\text{data}}(\cdot))$, with Y defined by (4.2). [14] first provided a bound of $d_{TV}(\mathcal{L}(Y_T), p_{\text{data}}(\cdot))$ for the OU process (3.1), assuming an L^∞-bound on score matching (Assumption 5.1'). The bound is of form:

$$d_{TV}(\mathcal{L}(Y_T), p_{\text{data}}(\cdot)) \leq \underbrace{d_{TV}(p(T, \cdot), p_{\text{noise}}(\cdot))}_{\leq C \exp(-C'T)} + D\varepsilon \exp(D'T), \tag{5.1}$$

for some $C, C', D, D' > 0$ depending on θ, μ and $p_{\text{data}}(\cdot)$. The disadvantage of this result is that the score matching error grows exponentially in time T. A recent breakthrough [10] improved this bound to be polynomial in T (and d). The result is stated for the (discretized) OU process, but is easily extended to the general case.

Theorem 5.2 (Total variation bound). [10] *Let Assumption 5.1 hold, and assume that $g(t)$ is bounded away from zero. Then*

$$d_{TV}(\mathcal{L}(Y_T), p_{data}(\cdot)) \leq \underbrace{d_{TV}(p(T, \cdot), p_{noise}(\cdot))}_{\text{noise approx. error}} + \underbrace{\varepsilon\sqrt{T/2}}_{\text{score matching error}}. \tag{5.2}$$

Proof. We sketch the proof. Denote by

- Q_T the distribution of the backward process (4.2), i.e., with the matched score $s_\theta(t, x)$ and $Y_0 \sim p_{\text{noise}}(\cdot)$;
- Q_T' the distribution of the backward process, with the matched score $s_\theta(t, x)$ and $Y_0 \sim p(T, \cdot)$;
- Q_T'' the distribution of the backward process, with the true score $\nabla \log p(t, x)$ and $Y_0 \sim p(T, \cdot)$.

By the data processing inequality, we have:

$$
\begin{aligned}
d_{TV}(\mathcal{L}(Y_T), p_{\text{data}}(\cdot)) &\leq d_{TV}(Q_T, Q_T') + d_{TV}(Q_T', Q_T'') \\
&\leq d_{TV}(p(T, \cdot), p_{\text{noise}}(\cdot)) + d_{TV}(Q_T', Q_T'').
\end{aligned}
\tag{5.3}
$$

The key to bound the term $d_{TV}(Q_T', Q_T'')$ is the identity:

$$
\mathrm{KL}(Q_T'', Q_T') = \mathbb{E}_{Q_T''} \left(\log \frac{dQ_T''}{dQ_T'} \right) = \mathbb{E}_{Q_T''} \int_0^T |s_\theta(T - t, Y_t) - \nabla p(T - t, Y_t)|^2 dt,
\tag{5.4}
$$

where the second equality follows from the Girsanov theorem. Further by Assumption 5.1, we get by Pinsker's inequality:

$$
d_{TV}(Q_T', Q_T'') \leq \sqrt{\frac{1}{2} \mathrm{KL}(Q_T'', Q_T')} \leq \varepsilon \sqrt{T/2}.
\tag{5.5}
$$

Combining (5.3) and (5.5) yields (5.2). $\qquad \square$

Now let's consider an important special case – the VP SDE that also includes the OU process.

Corollary 5.3 (Total variation for VP). *Let $(Y_t, 0 \leq t \leq T)$ be specified by (3.8) (the backward process of VP). Let Assumption 5.1 hold, and assume that $\beta(t)$ is bounded away from zero, and $\mathbb{E}_{p_{data}(\cdot)} |x|^2 < \infty$. Then for T sufficiently large,*

$$
d_{TV}(\mathcal{L}(Y_T), p_{data}(\cdot)) \leq e^{-\frac{1}{2} \int_0^T \beta(s) ds} \sqrt{\mathbb{E}_{p_{data}(\cdot)} |x|^2 / 2} + \varepsilon \sqrt{T/2}.
\tag{5.6}
$$

Proof. Recall from (3.10) that for VP, the conditional distribution of $(X_t \,|\, X_0 = x)$ is $\mathcal{N}\left(e^{-\frac{1}{2} \int_0^t \beta(s) ds} x, (1 - e^{-\int_0^t \beta(s) ds}) I\right)$. By the convexity of KL divergence, we have:

$$
\mathrm{KL}(p(T, \cdot), p_{\text{noise}}(\cdot)) \leq \mathbb{E}_{p_{\text{data}}(\cdot)} \left[\underbrace{\mathrm{KL}\left(\mathcal{N}\left(e^{-\frac{1}{2} \int_0^T \beta(s) ds} x, (1 - e^{-\int_0^T \beta(s) ds}) I\right), \mathcal{N}(0, I) \right)}_{(a)} \right].
\tag{5.7}
$$

Moreover,

$$
\begin{aligned}
(a) &= e^{-\int_0^T \beta(s) ds} |x|^2 - d \left(e^{-\int_0^T \beta(s) ds} + \log(1 - e^{-\int_0^T \beta(s) ds}) \right) \\
&\leq e^{-\int_0^T \beta(s) ds} |x|^2, \quad \text{as } T \to \infty,
\end{aligned}
\tag{5.8}
$$

since $\beta(t)$ is bounded away from zero, and hence $\int_0^T \beta(s) ds \to \infty$ as $T \to \infty$. Combining (5.7), (5.8) with (5.2) yields (5.6). $\qquad \square$

Similar polynomial bounds were also obtained by [40, 41] for the VP SDE. There χ^2-divergence (and an identity analogous to (5.4)) is used to bound $d_{TV}(\mathcal{L}(Y_T), p_{\text{data}}(\cdot))$.

5.2. **Wasserstein bound.** We make the following additional assumptions.

Assumption 5.4. *The following conditions hold:*

(1) *There exists $r_f : [0, T] \to \mathbb{R}$ such that $(x - x') \cdot (f(t, x) - f(t, x')) \geq r_f(t)|x - x'|^2$ for all t, x, x'.*

(2) *There exists $L > 0$ such that $|\nabla \log p(t, x) - \nabla \log p(t, x')| \leq L|x - x'|$ for all t, x, x'.*

The condition (1) assumes the growth of $f(t, \cdot)$, and (2) assumes the Lipschitz property of the score function. In all the examples of Section 3, $f(t, x)$ is linear in x so the density $p(t, \cdot)$ is Gaussian-like, and its score is almost affine. Thus, it is reasonable to assume (2). The following result gives a bound of $W_2(\mathcal{L}(Y_T), p_{\text{data}}(\cdot))$.

Theorem 5.5 (Wasserstein bound). [70] *Let Assumptions 5.1', 5.4 hold. For $h > 0$ a hyperparameter, define*

$$u(t) := \int_{T-t}^{T} \left(-2r_f(s) + (2L + 2h)g^2(s) \right) ds. \tag{5.9}$$

Then we have

$$W_2(p_{data}(\cdot), \mathcal{L}(Y_T)) \leq \sqrt{\underbrace{W_2^2(p(T, \cdot), p_{noise}(\cdot))e^{u(T)}}_{\text{noise approx. error}} + \underbrace{\frac{\varepsilon^2}{2h} \int_0^T g^2(t)e^{u(T)-u(T-t)}dt}_{\text{score matching error}}}. \tag{5.10}$$

Proof. The idea relies on coupling. Consider the coupled SDEs:

$$\begin{cases} dU_t = \left(-f(T - t, U_t) + g^2(T - t)\nabla \log p(T - t, U_t) \right) dt + g(T - t)dB_t, \\ dV_t = \left(-f(T - t, V_t) + g^2(T - t)s_\theta(T - t, V_t) \right) dt + g(T - t)dB_t, \end{cases}$$

where (U_0, V_0) are coupled to achieve $W_2(p(T, \cdot), p_{\text{noise}}(\cdot))$. Note that $W_2^2(p_{\text{data}}(\cdot), \mathcal{L}(Y_T)) \leq \mathbb{E}|U_T - V_T|^2$, so the goal is to bound $\mathbb{E}|U_T - V_T|^2$. By Itô's formula, we get

$$d|U_t - V_t|^2 = 2(U_t - V_t) \cdot (-f(T - t, U_t) + g^2(T - t)\nabla \log p(T - t, U_t)$$
$$+ f(T - t, V_t) - g^2(T - t)s_\theta(T - t, V_t))dt,$$

which implies that

$$\frac{d\,\mathbb{E}|U_t - V_t|^2}{dt} = -2\underbrace{\mathbb{E}((U_t - V_t) \cdot (f(T - t, U_t) - f(T - t, V_t)))}_{(a)}$$
$$+ 2\underbrace{\mathbb{E}((U_t - V_t) \cdot g^2(T - t)(\nabla \log p(T - t, U_t) - s_\theta(T - t, V_t)))}_{(b)}. \tag{5.11}$$

By Assumption 5.4 (1), we get $(a) \geq r_f(T - t)\,\mathbb{E}|U_t - V_t|^2$. Moreover,

$$(b) = g^2(T - t)\left(\mathbb{E}((U_t - V_t) \cdot (\nabla \log p(T - t, U_t) - \nabla \log p(T - t, V_t))) \right.$$

$$\left. + \mathbb{E}((U_t - V_t) \cdot (\nabla \log p(T - t, V_t) - s_\theta(T - t, V_t))) \right) \tag{5.12}$$

$$\leq g^2(T - t)\left(h\mathbb{E}|U_t - V_t|^2 + \frac{1}{4h}\varepsilon^2 + L\,\mathbb{E}|U_t - V_t|^2 \right),$$

where we use Assumptions 5.1' and 5.4 (3) in the last inequality. Therefore,

$$\frac{d\,\mathbb{E}|U_t - V_t|^2}{dt} \leq \left(-2r_f(T-t) + (2h+2L)g^2(T-t)\right)\mathbb{E}|U_t - V_t|^2 + \frac{\varepsilon^2}{2h}g^2(T-t).$$

Applying Grönwall's inequality, we have $\mathbb{E}|U_T - V_T|^2 \leq e^{u(T)}\mathbb{E}|U_0 - V_0|^2 + \frac{\varepsilon^2}{2h}\int_0^T \sigma^2(T - t)e^{u(T)-u(t)}dt$, which yields (5.10). $\qquad\square$

Theorem 5.5 assumes an L^∞-bound on score matching. Note that if an L^2-bound is assumed (Assumption 5.1), the result still holds if Assumption 5.4 (2) is replaced with

$$|s_\theta(t,x) - s_\theta(t,x')| \leq L|x - x'| \quad \text{for all } t, x, x'. \tag{5.13}$$

That is, the score matching function rather than the score function is Lipschitz, as considered in [39]. However, the condition (5.13) seems to be too restrictive, say for a neural net. [22] obtained a similar result by an L^2-bound on score matching but directly assuming that $\mathbb{E}|s_\theta(T - t, V_t) - \nabla \log p(T - t, V_t)|^2 \leq \varepsilon^2$ for all t.

Theorem 5.5 does not require a specific structure of $f(t,x)$. The bound (5.10) is loose in the general case, e.g. $r_f(t)$ may be negative, and hence the bound is exponential in T. Nevertheless, the estimates in the proof can be refined in some special cases. Here we examine two examples: VP SDE and CVP SDE.

The following result provides a bound for VP, assuming that $p_{\text{data}}(\cdot)$ is strongly log-concave. Recall that a smooth function $h : \mathbb{R}^d \to \mathbb{R}$ is κ-strongly log-concave if

$$(\nabla \log h(x) - \nabla \log h(y)) \cdot (x - y) \leq -\kappa|x - y|^2. \tag{5.14}$$

In comparison with Corollary 5.3, the score matching error does not grow in T.

Theorem 5.6 (Wasserstein bound for VP). *Let $(Y_t, 0 \leq t \leq T)$ be defined by (3.8) (the backward process of VP). Assume that $p_{data}(\cdot)$ is κ-strongly log-concave with $\kappa > \frac{1}{2}$, and $\mathbb{E}_{p_{data}(\cdot)}|x|^2 < \infty$. For $h < \min(\kappa, 1) - \frac{1}{2}$, we have:*

$$W_2^2(p_{data}(\cdot), \mathcal{L}(Y_T)) \leq e^{-\{\beta_{\min} + \beta_{\max}(2\min(\kappa,1) - 1 - 2h)\}T}\left(\mathbb{E}_{p_{data}(\cdot)}|x|^2 + o(e^{-\beta_{\min}T})d\right)$$
$$+ \varepsilon^2 \frac{\beta_{\max}^2}{2h(2\min(\kappa,1) - 1 - 2h)}. \tag{5.15}$$

Proof. Recall that $r_f(t) = -\frac{1}{2}\beta(t)$, $g(t) = \sqrt{\beta(t)}$ and $p_{\text{noise}}(\cdot) \sim \mathcal{N}(0, I)$. The key improvement relies on the fact [22, Proposition 10]: If $p_{\text{data}}(\cdot)$ is κ-strongly log-concave, then $\nabla \log p(T - t, \cdot)$ is $\kappa\left(e^{-\int_0^{T-t} \beta(s)ds} + \kappa\int_0^{T-t} e^{-\int_s^{T-t}\beta(v)dv}\beta(s)ds\right)^{-1}$-strongly concave. (The convolution preserves strongly log-concavity.) Thus, the term $\mathbb{E}((U_t - V_t) \cdot (\nabla \log p(T - t, U_t) - \nabla \log p(T - t, V_t)))$ in (5.12) is bounded from above by

$$-\frac{\kappa}{e^{-\int_0^{T-t}\beta(s)ds} + \kappa\int_0^{T-t}e^{-\int_s^{T-t}\beta(v)dv}\beta(s)ds}\mathbb{E}|U_t - V_t|^2.$$

instead of $L\,\mathbb{E}|U_t - V_t|^2$. Consequently, we obtain the bound (5.10) by replacing $u(t)$ with

$$u_{\text{VP}}(t) := \int_{T-t}^T \beta(s)\left(1 + 2h - \frac{2\kappa}{e^{-\int_0^s \beta(v)dv} + \kappa\int_0^s e^{-\int_v^s \beta(u)du}\beta(v)dv}\right)ds. \tag{5.16}$$

It is easily seen that $u_{\mathrm{VP}}(T) \leq \beta_{\max}(1 + 2h - 2\min(\kappa, 1))T$ and $u_{\mathrm{VP}}(T) - u_{\mathrm{VP}}(T - t) \leq \beta_{\max}(1 + 2h - 2\min(\kappa, 1))t$. Moreover,

$$W_2^2(p(T, \cdot), p_{\mathrm{noise}}(\cdot)) \leq e^{-\int_0^T \beta(s)ds}\mathbb{E}_{p_{\mathrm{data}(\cdot)}}|x|^2 + \frac{d}{4}e^{-2\int_0^T \beta(s)ds}$$

$$\leq e^{-\beta_{\min}T}\mathbb{E}_{p_{\mathrm{data}(\cdot)}}|x|^2 + \frac{d}{4}e^{-2\beta_{\min}T}.$$

For $\kappa > \frac{1}{2}$, we pick $h < \min(\kappa, 1) - \frac{1}{2}$, so $1 + 2h - 2\min(\kappa, 1) < 0$. Combining (5.10) with the above estimates yields (5.15). □

Next we give a bound for CVP. As we will see, the model parameters need to be tuned with respect to T.

Theorem 5.7 (Wasserstein bound for CVP). *Let $(Y_t, 0 \leq t \leq T)$ be specified by (3.18) (the backward process of CVP). Assume that $p_{data}(\cdot)$ is κ-strongly log-concave, and $\mathbb{E}_{p_{data}(\cdot)}|x|^2 < \infty$. We have:*

$$W_2^2(p_{data}(\cdot), \mathcal{L}(Y_T)) \leq e^{-2\left(\frac{\kappa}{1+\kappa} - \beta_{\max}hT + \mathcal{O}(e^{-\beta_{\min}T})\right)}\mathbb{E}_{p_{data}(\cdot)}|x|^2 + \frac{\varepsilon^2}{2h(1 - 2h)}. \tag{5.17}$$

Proof. Recall that $r_f(t) = \frac{1}{2}\beta(t)$, $g(t) = \sqrt{\beta(t)}$ and $p_{\mathrm{noise}}(\cdot) \sim \mathcal{N}(0, (e^{\frac{T}{2}(\beta_{\max}+\beta_{\min})} - 1)I)$. Similar to the proof of Theorem 5.6, the bound (5.10) holds by replacing $u(t)$ with

$$u_{\mathrm{CVP}}(t) := \int_{T-t}^T \beta(s)\left(-1 + 2h - \frac{2\kappa}{e^{\int_0^s \beta(v)dv} + \kappa \int_0^s e^{\int_v^s \beta(u)du}\beta(v)dv}\right)ds. \tag{5.18}$$

Note that $u_{\mathrm{CVP}}(T) \leq -\int_0^T \beta(s)ds + 2\beta_{\max}hT - \frac{2\kappa}{\kappa+1}\left(1 - e^{-\beta_{\min}T}\right)$ and $u_{\mathrm{CVP}}(T) - u_{\mathrm{CVP}}(T-t) \leq \beta_{\max}(2h-1)t$. Moreover, $W_2^2(p(T, \cdot), p_{\mathrm{noise}}(\cdot)) \leq e^{\int_0^T \beta(s)ds}\mathbb{E}_{p_{\mathrm{data}}(\cdot)}|x|^2$. Combining (5.10) with the above estimates yields (5.17). □

Let's make several comments. First, comparing Theorem 5.6 with Theorem 5.7, we see that for VP both the noise approximation and the score matching error depend on the strong log-concavity of $p_{\mathrm{data}}(\cdot)$, whereas for CVP the score matching error does not depend on the latter. For the bound (5.17), we typically take

$$\beta_{\min}T \to \infty \quad \text{and} \quad \beta_{\max}hT < \frac{\kappa}{\kappa + 1}, \tag{5.19}$$

so the factor $e^{-2\left(\frac{\kappa}{1+\kappa} - \beta_{\max}hT + \mathcal{O}(e^{-\beta_{\min}T})\right)} < 1$. This implies that CVP may incurs a noise approximation error of constant order (see Table 1 for a summary).

	Noise approximation error	Score matching error
VP	Exponential decay in T	Depend on strongly log-concavity of $p_{\mathrm{data}}(\cdot)$
CVP	Bounded	Not depend on strongly log-concavity of $p_{\mathrm{data}}(\cdot)$

TABLE 1. Comparison of VP and CVP in W_2 metric.

The assumption (5.19) suggests that h scale inversely to $\beta_{\max}T \to \infty$. For instance, if we take $\beta_{\max}T = T^\gamma$ for some (small) $\gamma < 1$, then h is of order $T^{-\gamma}$, and the score matching error is of order $\varepsilon T^{\frac{\gamma}{2}}$. Thus, VP beats CVP at the *continuous* level by assuming the *strongly log-concavity* on $p_{\mathrm{data}}(\cdot)$. However, the target distribution is rarely strongly log-concave. Thus,

it is not clear how rapid the score matching error grows in T for VP. CVP (and more general CDPM) is designed to be robust to the score matching error, and to the discretization (see Section 5.3). Practically, T should not set to be too large for CVP to reduce the noise approximation bias.

5.3. **Discretization.** In order to implement the diffusion model, we need to discretize the corresponding SDE. For the SDE (4.2), there are several ways of discretization. Fix $0 = t_0 < t_1 \cdots < t_N = T$ to be the time discretization. Initialize $\widehat{X}_0 \sim p_{\text{noise}}(\cdot)$, and set

$$\widehat{X}_k = \widehat{X}_{k-1} + (-f(T-t_k, \widehat{X}_{k-1}) + a(T-t_{k-1})s_\theta(T - t_{k-1}, \widehat{X}_{k-1}))(t_k - t_{k-1})$$
$$+ \int_{t_{k-1}}^{t_k} g(T-t)dB_t, \quad \text{for } k = 1, \ldots, N, \tag{5.20}$$

where $\int_{t_{k-1}}^{t_k} g(T-t)dB_t$ is sampled as $\mathcal{N}(0, \int_{t_{k-1}}^{t_k} g^2(T-t)dt)$, or the Euler-Maruyama scheme:

$$\widehat{X}_k = \widehat{X}_{k-1} + (-f(T-t_k, \widehat{X}_{k-1}) + a(T - t_{k-1})s_\theta(T - t_{k-1}, \widehat{X}_{k-1}))(t_k - t_{k-1})$$
$$+ g(T - t_{k-1})(B_{t_k} - B_{t_{k-1}}), \quad \text{for } k = 1, \ldots, N. \tag{5.21}$$

In practice, the predictor-corrector (PC) sampler [67] is frequently used, where the scheme (5.20) or (5.21) serves as the predictor, followed by a score-based corrector. More precisely, for each k, let $\widehat{X}_k^{(0)}$ be the output of the predictor (5.20) or (5.21), and fix some $\epsilon_k > 0$. The corrector $\widehat{X}_k = \widehat{X}_k^{(M)}$ is given by

$$\widehat{X}_k^{(\ell+1)} = \widehat{X}_k^{(\ell)} + \epsilon_k s_\theta(T - t_k, \widehat{X}_k^{(\ell)}) + \sqrt{2}(B'_{(\ell+1)\epsilon_k} - B'_{\ell\epsilon_k}), \quad \text{for } \ell = 0, \ldots, M - 1, \tag{5.22}$$

where M is the number of steps in the corrector step. The hyperparameters ϵ_k are taken to be decreasing in k. Essentially, (5.22) is the discrete scheme of the Langevin equation $dX_t = \nabla \log p(T - t_k, X_t) + \sqrt{2}B'_t$ that converges to $p(T - t_k, \cdot)$, but with the score function replaced with the estimated $s_\theta(T - t_k, \cdot)$. Since the scheme (5.20) or (5.21) is the predictor of the PC sampler, it is also called the predictor-only sampler.

Classical SDE theory [37] indicates that the discretization error is of order $C(T)\delta$, with $C(T)$ exponential in T. A natural question is whether there are better bounds for the SDEs arising from the diffusion models in Section 3. The general way to control the discretization error is to prove a one-step estimate, and then unfold the recursion. The analysis is mostly tedious, and we do not plan to expand the details here. The table below summarizes known results on the discretization error of various diffusion models.

Reference	Evaluation metric	Model	Sampling	Step size	Discretization error
[10]	TV	OU	Euler	δ	$\sqrt{T}d\sqrt{\delta}$
[40, 41]	TV	OU	PC	varying	$\mathcal{O}_T(1)\sqrt{d}\sqrt{\delta}$
[22]	W_2	VE	Euler	δ	$e^{\mathcal{O}(T)}\sqrt{d}\delta$
	W_2	VP	Euler	δ	$e^{\mathcal{O}(T^3)}\sqrt{d}\delta$
[70]	W_2	CDPM	Euler	δ	$\mathcal{O}_T(1)\sqrt{d}\sqrt{\delta}$

TABLE 2. Discretization errors.

There has been work on the convergence of diffusion models in a compact domain, or satisfying the manifold hypothesis [13, 53]. In this case, the Wasserstein distance is simply

bounded by the total variation distance. [10, 40, 41] also considered bounded target distributions, which should not be confused with diffusion models in a compact domain because the SDEs can live in an unbounded space. There an early stopping argument is used to guarantee the convergence of diffusion models in the W_2 metric. See also [2] for finer results on the convergence of the OU process, and [43, 45] that worked directly on the convergence of (discrete) DDPM.

6. ODE samplers and consistency model

The stochastic sampler (4.2) and its discretization (5.20) may suffer from slow convergence and instability due to the randomness. In this section, we consider sampling Y_T by an equivalent deterministic scheme – the ODE sampler. Section 6.1 presents the probability flow ODE that leads to the ODE sampler of diffusion models. In Section 6.2, we discuss one prominent recent application of ODE sampling – the consistency model.

6.1. **The probability flow ODE.** The main idea is that starting from the noise $Y_0' \sim p_{\text{noise}}$, and going through an ODE rather than an SDE will produce a sample Y_T' that has the same distribution as Y_T. Here we are only concerned with the sampled distribution at time T.

The equivalent ODE formulation is a consequence of the following result, known as the *probability flow ODE*.

Theorem 6.1 (Probability flow ODE). *Let $(X_t, 0 \le t \le T)$ be defined by (2.1), with $p(t, \cdot)$ the probability distribution of X_t. Let*

$$\widetilde{f}(t, x) := f(t, x) - \frac{1}{2} \nabla \cdot (g(t, x)g(t, x)^\top) - \frac{1}{2} g(t, x)g(t, x)^\top \nabla \log p(t, x), \qquad (6.1)$$

and let $(\widetilde{X}_t, 0 \le t \le T)$ solve the ODE:

$$\frac{d\widetilde{X}_t}{dt} = \widetilde{f}(t, \widetilde{X}_t), \quad \widetilde{X}_0 \sim p_{data}(\cdot). \qquad (6.2)$$

Then for each t, X_t and $\widetilde{X}_t$ have the same distribution.

Proof. The idea is to match the Fokker-Planck equation in both settings. Denote by $\widetilde{p}(t, x)$ the probability distribution of $\widetilde{X}_t$ defined by (6.2). By the (first-order) Fokker-Planck equation, $\widetilde{p}(t, x)$ solves

$$\frac{\partial \widetilde{p}(t, x)}{\partial t} = -\nabla \cdot \left(\widetilde{f}(t, x)\widetilde{p}(t, x) \right), \quad \widetilde{p}(0, x) = p_{\text{data}}(x). \qquad (6.3)$$

On the other hand, $p(t, x)$ solves the (second order) Fokker-Planck equation:

$$\frac{\partial p(t, x)}{\partial t} = -\nabla \cdot (f(t, x)p(t, x)) + \frac{1}{2} \nabla^2 : \left(g(t, x)g(t, x)^\top p(t, x) \right), \quad p(0, x) = p_{\text{data}}(x). \quad (6.4)$$

where : denotes the Frobenius inner product. By algebraic rearrangement, (6.4) simplifies to:

$$\frac{\partial p(t, x)}{\partial t} = -\nabla \cdot \left(\widetilde{f}(t, x)p(t, x) \right), \quad p(0, x) = p_{\text{init}}(x). \qquad (6.5)$$

Comparing (6.2) with (6.4) yields $\widetilde{p}(t, \cdot) = p(t, \cdot)$. $\qquad \square$

We remark that the two processes X and $\widetilde{X}$ only have the same marginal distributions, but they don't have the same distribution at the level of the process (nor the joint distributions).

Now let's apply the equivalent ODE formulation to the diffusion models. We specify to the practical case where $n = d$ and $\sigma(t, x) = \sigma(t)I$. The corresponding ODE is given by

$$\frac{dX_t}{dt} = f(t, X_t) - \frac{1}{2}g^2(t)\nabla p(t, X_t), \quad X_0 \sim p_{\text{data}}(\cdot), \tag{6.6}$$

and the backward ODE is then:

$$\frac{dY_t}{dt} = -f(T - t, Y_t) + \frac{1}{2}g^2(T - t)\nabla p(T - t, Y_t), \quad Y_0 \sim p_{\text{noise}}(\cdot). \tag{6.7}$$

With the score function $\nabla \log p(t, x)$ being replaced with the score matching function $s_\theta(t, x)$, the backward ODE becomes:

$$\frac{dY_t}{dt} = -f(T - t, Y_t) + \frac{1}{2}g^2(T - t)s_\theta(T - t, Y_t), \quad Y_0 \sim p_{\text{noise}}(\cdot), \tag{6.8}$$

which can be regarded as the continuum limit of *denoising diffusion implicit models* (DDIM) [61].

(a) The ODE sampler for OU is:

$$\frac{dY_t}{dt} = \theta(Y_t - \mu) + \frac{1}{2}\sigma^2 s_\theta(T - t, Y_t), \quad Y_0 \sim p_{\text{noise}}(\cdot). \tag{6.9}$$

(b) The ODE sampler for VE is:

$$\frac{dY_t}{dt} = \frac{1}{2}g^2(T - t)s_\theta(T - t, Y_t), \quad Y_0 \sim p_{\text{noise}}(\cdot), \tag{6.10}$$

where g is defined by (3.5). As seen in the next subsection, practitioners also take $g(t) = \sqrt{2t}$ for the consistency model.

(c) The ODE sampler for VP is:

$$\frac{dY_t}{dt} = \frac{1}{2}\beta(T - t)\left(Y_t + s_\theta(T - t, Y_t)\right), \quad Y_0 \sim p_{\text{noise}}(\cdot). \tag{6.11}$$

where $\beta(t)$ is defined by (3.9).

(d) The ODE sampler for sub-VP is:

$$\frac{dY_t}{dt} = \frac{1}{2}\beta(T - t)\left(Y_t + (1 - \gamma(T - t))s_\theta(T - t, Y_t)\right), \quad Y_0 \sim p_{\text{noise}}(\cdot). \tag{6.12}$$

(e1) The ODE sampler for COU is

$$\frac{dY_t}{dt} = -\theta(Y_t - \mu) + \frac{1}{2}\sigma^2 s_\theta(T - t, Y_t), \quad Y_0 \sim p_{\text{noise}}(\cdot). \tag{6.13}$$

(e2) The ODE sampler for CVP is

$$\frac{dY_t}{dt} = -\frac{1}{2}\beta(T - t)\left(Y_t - s_\theta(T - t, Y_t)\right), \quad Y_0 \sim p_{\text{noise}}(\cdot). \tag{6.14}$$

The ODEs (6.9)–(6.14) can be easily solved by existing ODE solvers. The theory of the ODE sampling of diffusion models is, however, still in its infancy. [9] studied the total variation convergence of the ODE sampler for the OU process, and [23] established Wasserstein

bounds for the ODE sampler of VE and VP. See also [3, 11, 43, 45] for closely related results on the deterministic sampler of diffusion models.

Note that in all the above examples, $f(\cdot, \cdot)$ takes the form $f(t, x) = f(t)x$ with $f(t) \in \mathbb{R}_+$. In this case, by a change of variables [35], the diffusion model can be reparametrized by

$$s(t) := \exp\left(\int_0^t f(r)dr\right) \quad \text{and} \quad \ell(t) := \sqrt{\int_0^t \frac{g^2(r)}{s(r)}dr}, \tag{6.15}$$

so that the ODE (6.6) becomes:

$$\frac{dX_t}{dt} = \frac{s'(t)}{s(t)}x - s(t)^2 \ell'(t)\ell(t)\nabla \log p\left(\ell(t), \frac{x}{s(t)}\right), \qquad X_0 \sim p_{\text{data}}(\cdot). \tag{6.16}$$

Empirical results show that the choice of $s(t) = 1$ and $\ell(t) = t$ (i.e., VE with $f(t) = 0$ and $g(t) = \sqrt{2t}$) yields the state of the art.

6.2. **Consistency model.** Here we present an application of the probability flow ODE – the consistency model [63] (see [36, 62] for a more advanced version, and [48] for its latest application.) The advantage of the consistency modeling is that it only performs a *one-step* sampling – this is in contrast with the SDE or the ODE sampling, where multiple steps are required in discretization.

The main takeaway of the consistency model is to approximate the ODE flow. Recall the ODE sampler from (6.7). The ODE theory implies that there is a function (the ODE flow) $F : \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^d$ such that

$$Y_t = F(t, s, Y_s), \quad \text{for } 0 \le s < t \le T. \tag{6.17}$$

Thus, $Y_T = F(T, t, Y_t)$ for any t. The consistency model proposes to learn $F(T, \cdot, \cdot)$ directly. To be more precise, the function of form $F_\theta(T-t, y)$ is used to approximate the flow $F(T, t, y)$. (Note that if the ODE (6.7) is autonomous, e.g. the OU process, the flow $F(t, s, y)$ is indeed of form $F(t - s, y)$.) With the learned flow F_θ, the model outputs the sample $F_\theta(T, Y_0)$, $Y_0 \sim p_{\text{noise}}(\cdot)$ in just one step.

Now the question is how to learn the ODE flow $F(T, t, y) \approx F_\theta(T - t, y)$. Two methods were proposed based on the group property (6.17).

(a) *Consistency distillation* (CD): the approach builds on top of a pretrained score function $s_\theta(t, x)$. Take $t \sim \mathcal{U}(0, T)$. First sample $Y_+ \overset{d}{=} X_t$ of the forward process (VE with $g(t) = \sqrt{2t}$ is used, so $X_t \sim \mathcal{N}(x, t^2 I)$ with $x \sim p_{\text{data}}(\cdot)$). Then use the ODE sampler (6.7) to set:

$$Y_- = Y_+ + \left(-f(t, Y_+) + \frac{1}{2}g^2(t)s_\theta(t, Y_+)\right)\delta, \tag{6.18}$$

for some small δ. Roughly speaking, (Y_+, Y_-) are on the ODE flow F at the times $T - t$ and $T - t + \delta$. Inspired by the fact that

$$\underbrace{F(T, T - t, Y_{T-t})}_{\approx F_\theta(t, Y_+)} = Y_T = \underbrace{F(T, T - t + \delta, Y_{T-t+\delta})}_{\approx F_\theta(t-\delta, Y_-)},$$

the training objective is:

$$\mathbb{E}_{t \in \mathcal{U}(0,T)}\mathbb{E}[\lambda(t)|F_\theta(t, Y_+) - F_\theta(t - \delta, Y_-)|^2], \tag{6.19}$$

where $\lambda(t)$ is a weight function.

(b) *Consistency training* (CT): the approach is similar to CD, except that $Y_- \stackrel{d}{=} X_{t-\delta}$ is taken *independently* of $Y_+ \stackrel{d}{=} X_t$. (In the VE setting with $g(t) = \sqrt{2t}$, $Y_+ \sim \mathcal{N}(x, t^2 I)$ and $Y_- \sim \mathcal{N}(x, (t-\delta)^2 I)$ independently.) Thus, CT does not require a pretrained score function. Note that

$$Y_+ \stackrel{d}{=} Y_{T-t} \quad \text{and} \quad Y_- \stackrel{d}{=} Y_{T-t+\delta} \quad \text{marginally.}$$

But contrary to CD, (Y_+, Y_-) are not on the ODE flow path, i.e. a wrong coupling.

Now we consider a simple scenario, where $g(t) = \sqrt{2t}$ and $p_{\text{data}}(\cdot) = \delta_0$ (delta mass at 0). It is easy to compute that

$$(Y_+^{CT}, Y_-^{CT}) \stackrel{d}{=} \mathcal{N}\left(0, \begin{pmatrix} t^2 I & 0 \\ 0 & (t-\delta)^2 I \end{pmatrix}\right), \quad (Y_+^{CD}, Y_-^{CD}) \stackrel{d}{=} \mathcal{N}\left(0, \begin{pmatrix} t^2 I & t(t-\delta)I \\ t(t-\delta)I & (t-\delta)^2 I \end{pmatrix}\right).$$

As a result,

$$\begin{aligned} W_2^2((Y_+^{CD}, Y_-^{CD}), (Y_+^{CT}, Y_-^{CT})) &= 2d\left(t^2 + (t-\delta)^2 - \sqrt{t^4 + (t-\delta)^4}\right) \\ &\approx 2(2 - \sqrt{2})t^2 d. \end{aligned} \tag{6.20}$$

So the distance between (Y_+^{CD}, Y_-^{CD}) and (Y_+^{CT}, Y_-^{CT}) is far from being negligible. By taking $t \sim \mathcal{U}(0,1)$, the expected W_2 distance is approximately $\sqrt{\left(1 - \frac{\sqrt{2}}{2}\right)d}$. This explains why CD outperforms CT empirically [63]. A remedy to CT has recently been proposed in [62].

Also note that the objective function (6.19) can be lifted to continuous time via partial differential relation of the ODE flow $Y_T = F(T, T-t, Y_{T-t}) \approx F_\theta(t, Y_{T-t})$:

$$\mathbb{E}_{t \in \mathcal{U}(0,T)} \mathbb{E}\left[\lambda(t) \left|\partial_t F_\theta(t, Y_+) + \nabla F_\theta(t, Y_+)\left(f(t, Y_+) - \frac{1}{2}g^2(t) s_\theta(t, Y_+)\right)\right|^2\right], \tag{6.21}$$

or

$$\mathbb{E}_{t \in \mathcal{U}(0,T)} \mathbb{E}\left[\lambda(t) F_\theta(t, Y_+) \cdot \left(\partial_t F_\theta(t, Y_+) + \nabla F_\theta(t, Y_+)\frac{Y_+ - x}{t}\right)\right], \tag{6.22}$$

where $Y_+ \stackrel{d}{=} X_t$ and $x \sim p_{\text{data}}(\cdot)$. The functions (6.21) and (6.22) are the continuum limits of CD and CT respectively after suitable scaling.

7. FURTHER RESULTS ON SCORE MATCHING

In the previous sections, we have seen that the success of diffusion models relies largely on the quality of score matching. This section reviews further results on score matching in the context of diffusion models. Section 7.1 considers statistical efficiency of score matching, and Section 7.2 presents approximations by neural nets. In Section 7.3, we introduce the idea of diffusion model alignment by reinforcement learning (RL).

7.1. **Statistical efficiency of score matching.** In this subsection, we focus on the statistical aspect of the score matching problem (4.4). Assume that

$$s_\theta(t, x) = \nabla \log p_\theta(t, x), \quad \text{for all } t, x, \tag{7.1}$$

where $p_\theta(t, \cdot)$ is a family of probability distributions parametrized by θ. In other words, $s_\theta(\cdot, \cdot)$ is the score function of a parametric family of probability distributions instead of an arbitrary neural net. Thus, the score matching problem (4.4) becomes

$$\min_\theta \mathcal{J}_{\mathrm{ESM}}(\theta) := \mathbb{E}_{p(t, \cdot)} |\nabla \log p_\theta(t, x) - \nabla \log p(t, X)|^2, \tag{7.2}$$

and the corresponding implicit score matching problem is:

$$
\begin{aligned}
\min_\theta \mathcal{J}_{\mathrm{ISM}}(\theta) :&= \mathbb{E}_{p(t, \cdot)} \left[|\nabla \log p_\theta(t, X)|^2 + 2 \nabla \cdot \nabla \log p_\theta(t, X) \right] \\
&= \mathbb{E}_{p(t, \cdot)} \left[|\nabla \log p_\theta(t, X)|^2 + 2 \operatorname{Tr}(\nabla^2 \log p_\theta(t, X)) \right].
\end{aligned}
\tag{7.3}
$$

Let $\widehat{\theta}_n$ be minimum point of the empirical version of (7.2) or (7.3) from n samples $X_1, \ldots, X_n \sim p(t, \cdot)$:

$$\widehat{\theta}_n := \arg\min_\theta \frac{1}{n} \sum_{i=1}^n \left(|\nabla \log p_\theta(t, X_i)|^2 + 2 \operatorname{Tr}(\nabla^2 \log p_\theta(t, X_i)) \right). \tag{7.4}$$

The question of interest is whether the estimated $p_{\widehat{\theta}_n}(t, \cdot)$ approximates well the probability distribution $p(t, \cdot)$. The answer is hinted in the following theorem.

Theorem 7.1 (Consistency and asymptotic normality). [31, 38] *Fix $t \in [0, T]$.*

(1) *Assume that $p(t, \cdot)$ has full support, and $p(t, \cdot) = p_{\theta_*}(t, \cdot)$ for some θ_* and $p(t, \cdot) \neq p_\theta(t, \cdot)$ for any $\theta \neq \theta_*$. Then $\mathcal{J}_{ESM}(\theta) = 0$ if and only if $\theta = \theta_*$.*

(2) *Assume further that $p_\theta(t, \cdot)$ belongs to the exponential family, i.e. $p_\theta(t, \cdot) \propto \exp(\theta \cdot F(\cdot))$ for a suitably nice F. Then*

$$\sqrt{n}(\widehat{\theta}_n - \theta_*) \to \mathcal{N}(0, \Gamma_{SM}), \tag{7.5}$$

with $\Gamma_{SM} := \mathbb{E}_{p_{\theta_*}(t, \cdot)} \left[(\nabla F(X) \nabla F(X)^\top)^{-1} \Sigma_{\nabla F(X) \nabla F(X)^\top \theta_* + \Delta F(X)} \mathbb{E}(\nabla F(X) \nabla F(X)^\top)^{-1} \right]$, where $\Sigma_\bullet$ is the Fisher information matrix, and the Laplacian $\Delta F(X)$ applies coordinate-wise to $F(X)$.

Proof. (1) If $\mathcal{J}_{\mathrm{ESM}}(\theta) = 0$, then $\nabla \log p_\theta(t, \cdot) = \nabla \log p(t, \cdot)$. So $p_\theta(t, \cdot) = p(t, \cdot) + C$, and $C = 0$ because $p(t, \cdot)$ and $p_\theta(t, \cdot)$ are probability distributions. By the hypothesis, we have $\theta = \theta_*$.

(2) Recall that $p(t, \cdot) = p_{\theta_*}(t, \cdot)$. The key to the analysis relies on the fact that

$$\theta_* = -\mathbb{E}(\nabla F(X) \nabla F(X)^\top)^{-1} \mathbb{E} \Delta F(X). \tag{7.6}$$

Denote by $\widehat{\mathbb{E}}_n$ the empirical expectation of n i.i.d. random variables distributed by $p(t, \cdot)$. Write $\widehat{\mathbb{E}}_n(\nabla F(X) \nabla F(X)^\top) = \mathbb{E}(\nabla F(X) \nabla F(X)^\top) + \delta_{1,n}/\sqrt{n}$ and $\widehat{\mathbb{E}}_n \Delta F(X) = \mathbb{E} F(X) +$

$\delta_2/\sqrt{n}$. Note that

$$\widehat{\theta} = -\widehat{\mathbb{E}}_n(\nabla F(X)\nabla F(X)^\top)^{-1}\widehat{\mathbb{E}}_n \Delta F(X)$$

$$= -\left(\underbrace{\mathbb{E}(\nabla F(X)\nabla F(X)^\top)^{-1}\widehat{\mathbb{E}}_n(\nabla F(X)\nabla F(X)^\top)}_{I+\mathbb{E}(\nabla F(X)\nabla F(X)^\top)^{-1}\delta_{1,n}/\sqrt{n}}\right)^{-1}\underbrace{\mathbb{E}(\nabla F(X)\nabla F(X)^\top)^{-1}\widehat{\mathbb{E}}_n \Delta F(X)}_{-\theta_*+\mathbb{E}(\nabla F(X)\nabla F(X)^\top)^{-1}\delta_{2,n}/\sqrt{n}}$$

$$= \left(-I + \mathbb{E}(\nabla F(X)\nabla F(X)^\top)^{-1}\frac{\delta_{1,n}}{\sqrt{n}} + \mathcal{O}\left(\frac{1}{n}\right)\right)\left(-\theta_* + \mathbb{E}(\nabla F(X)\nabla F(X)^\top)^{-1}\frac{\delta_{2,n}}{\sqrt{n}}\right)$$

$$= \theta_* - \mathbb{E}(\nabla F(X)\nabla F(X)^\top)^{-1}\frac{\theta_*\delta_{1,n}+\delta_{2,n}}{\sqrt{n}} + \mathcal{O}(1/\sqrt{n}),$$

where we use Slutsky's theorem in the third equation. Moreover,

$$\frac{\theta_*\delta_{1,n}+\delta_{2,n}}{\sqrt{n}} = \mathbb{E}(\nabla F(X)\nabla F(X)^\top\theta_* + \Delta F(X)) - \widehat{\mathbb{E}}_n(\nabla F(X)\nabla F(X)^\top\theta_* + \Delta F(X)),$$

which converges in distribution to $\mathcal{N}(0, \Sigma_{\nabla F(X)\nabla F(X)^\top\theta_*+\Delta F(X)})$. Combining the above estimates yields (7.5). $\qquad\square$

Under the assumptions in Theorem 7.1, it is clear that Assumption 5.1 (L^2-bound on score matching) holds if

$$\mathbb{E}_{p(t,\cdot)}\left(\sup_\theta |\partial_\theta p_\theta(t,X)|^{2+\epsilon}\right) < \infty,$$

for any $\epsilon > 0$. Another interesting result from [38] shows that under an isoperimetric assumption on $p(t,\cdot)$, the statistical efficiency of score matching is close to the maximum likelihood estimate by comparing Γ_{SM} with $\Gamma_{\mathrm{MLE}} = \Sigma_{F(X)}^{-1}$. Similar results were obtained by [52] for a more general class of distributions.

7.2. **Score matching by neural nets.** Here we consider score matching for VP. Recall $\gamma(t) := \exp\left(-2\int_0^t \beta(s)ds\right)$. By Tweedie's formula [17, 57], the score function can be written as:

$$\nabla \log p(t,x) = \frac{\gamma(t)^{\frac{1}{4}}}{1-\sqrt{\gamma(t)}}\mathbb{E}(X_0 \mid X_t = x) - \frac{1}{1-\sqrt{\gamma(t)}}x, \qquad (7.7)$$

Inspired by the special structure (7.7), it is natural to take the function approximation:

$$s_\theta(t,x) = \frac{\gamma(t)^{\frac{1}{4}}}{1-\sqrt{\gamma(t)}}f_\theta(t,x) - \frac{1}{1-\sqrt{\gamma(t)}}x, \qquad (7.8)$$

with $f_\theta(t,x)$ a suitable neural net (e.g. ReLU net). As pointed out by [7], the specific form (7.8) has close affinity to the U-net, where the term $-\frac{1}{1-\sqrt{\gamma(t)}}x$ corresponds to a shortcut connection.

The following theorem provides a template for a class of results concerning the approximation efficiency of score matching. To simplify the presentation, we take the weight $\lambda(t) = 1$.

Theorem 7.2. [7, 25] *Let $\widehat{s}_\theta(\cdot, \cdot)$ solve one of the empirical score matching problems (in Section 4), where $\mathbb{E}$ is replaced with $\widehat{\mathbb{E}}$ under finite samples. Let $\widehat{s}_{\theta,k}$ be the output at iteration k of some algorithm (e.g. SGD). For $\widehat{s} = \widehat{s}_\theta$ or $\widehat{s}_{\theta,k}$, we have:*

$$\int_0^T \mathbb{E}_{p(t,\cdot)} |\widehat{s}(t,\cdot) - \nabla \log p(t,\cdot)|^2 dt$$

$$\leq \int_0^T \mathbb{E}_{p(t,\cdot)} |\widehat{s}(t,\cdot) - \nabla \log p(t,\cdot)|^2 - \widehat{\mathbb{E}}_{p(t,\cdot)} |\widehat{s}(t,\cdot) - \nabla \log p(t,\cdot)|^2 dt \qquad (7.9)$$

$$+ \inf_\theta \int_0^T \widehat{\mathbb{E}}_{p(t,\cdot)} |s_\theta(t,\cdot) - \nabla \log p(t,\cdot)|^2 dt,$$

where the first term on the right side of (7.9) *is the sample error of score matching, and the second term is the approximation error.*

In [7], $f_\theta(t,x)$ is taken to be a multilayer ReLU with linear structure. But no specific algorithm of stochastic optimization is used (corresponding to $\widehat{s} = \widehat{s}_\theta$). [25] considered a two-layer ReLU net, and gradient descent is used to update $\widehat{s} = \widehat{s}_{\theta,k}$. In particular, they used a neural tangent kernel as a universal proxy to the score, and showed that score matching (using a two-layer ReLU net) is as easy as kernel regression. However, the bound (7.9) depends, in a complicated manner, on the data distribution, the net structure, the sample size, and the number of iterations in the optimization algorithm.

7.3. **Score matching by reinforcement learning.** In this last subsection, we consider score matching by RL, which was initiated by [4, 18, 42]. The idea is to treat the score as the action in RL, where the reward measures how well the sample Y_T approximates $p_{\text{data}}(\cdot)$, either by a predetermined loss function, or by human feedback (diffusion model alignment). Since the diffusion model is specified by the SDE, the recently developed continuous RL [32, 33, 74, 77] provides a convenient framework to formulate score matching by RL.

Let's recall the setup of continuous RL. The state dynamics $(X_t^a, t \geq 0)$ is governed by the SDE:

$$dX_t^a = b(t, X_t^a, a_t) dt + \sigma(t) dB_t, \quad X_0^a \sim p_{\text{init}}(\cdot), \qquad (7.10)$$

where $b : [0,T] \times \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}^d$, $\sigma : [0,T] \to \mathbb{R}_+$, and the action $a_s \in \mathcal{A}$ is generated from a distribution $\pi(\cdot \,|\, t, X_t^a)$ by external randomization. Define the (state) value function given the feedback policy $\{\pi(\cdot \,|\, t, x) : x \in \mathbb{R}^d\}$ by

$$V(t, x; \pi) := \mathbb{E}\left[\int_t^T e^{-\beta(s-t)} r(s, X_s^\pi, a_s^\pi) \, ds \,\middle|\, X_0^\pi = x \right],$$

and the performance metric $\eta(\pi) := \int_{\mathbb{R}^d} V(0, x; \pi) p_{\text{init}}(dx)$. The main task of continuous RL is to approximate $\max_\pi \eta(\pi)$ by constructing a sequence of policies π_k, $k = 1, 2, \ldots$ recursively such that $\eta(\pi_k)$ is non-decreasing.

Now we introduce the main idea from [78]. Comparing the backward process (2.7) with the continuous RL (7.10), it is natural to set $p_{\text{init}}(\cdot) = p_{\text{noise}}(\cdot)$,

$$b(t, x, a) = -f(T - t, x) + g^2(T - t)a \quad \text{and} \quad \sigma(t) = g(T - t). \qquad (7.11)$$

Here the action a has a clear meaning – it is at the place of the score function. So we can deploy the continuous RL machinery for score matching. At time t, we use a Gaussian

exploration $a_t \sim \pi_\theta(\cdot \mid t, X_t) = N(\mu_\theta(t, X_t), \sigma_t^2 I)$, where σ_t^2 is a predetermined exploration level for each t. Under the Gaussian exploration, the RL dynamics is:

$$dX_t^\theta = \left[-f(T - t, X_t^\theta) + g^2(T - t)\mu_\theta(t, X_t^\theta)\right] dt + g(T - t)dB_t, \quad X_0^\theta \sim p_{\text{noise}}(\cdot). \quad (7.12)$$

Denote by $p_\theta(t, \cdot)$ the probability distribution of X_t^θ. In particular, the score matching parametrization $\mu_\theta(t, X_t^\theta) = s_\theta(T - t, X_t^\theta)$ recovers the backward process (4.2).

Next the reward is given by $r(t, x, a)$. In order to prevent the model from overfitting to the reward or catastrophic forgetting, we add a penalty term to ensure that the pretrained model $s_*(\cdot, \cdot)$ is fully utilized. The regularized reward is:

$$r_R(t, x, a) = e^{\alpha t} \left(r(t, x, a) + \gamma(t)|a - s_*(t, x)|^2\right), \quad (7.13)$$

for some $\alpha > 0$ and weight function $\gamma(t)$. The L^2 regularization is derived in [78] as an upper bound of the KL divergence between the pretrained model and the distribution generated by the current policy. The factor $e^{\alpha t}$, which corresponds to $\beta = -\alpha < 0$ in the value function (7.3), weighs the importance to the later rewards. For instance, if $\alpha \to \infty$, then only the final reward is taken into account.

The following theorem gives the policy gradient with respect to the performance metric η.

Theorem 7.3 (Policy gradient for score matching). [78] *Given a pretrained score s_*, the policy gradient of a policy π^θ parameterized by θ is:*

$$\nabla_\theta \eta(\pi_\theta) = \mathbb{E}\left[\int_0^T e^{\alpha t}\nabla_\theta \log \pi_\theta(a_t^{\pi_\theta}|t, X_t^\theta)\left(r(t, X_t^\theta, a_t^{\pi_\theta}) + b(t, X_t^\theta, a_t^{\pi_\theta}) \cdot \frac{\partial V}{\partial x}\left(t, X_t^\theta; \pi_\theta\right)\right.\right.$$
$$\left.\left. + \gamma(t) \cdot |a_t^{\pi_\theta} - s_*(t, X_t^\theta)|^2\right)dt\right]. \quad (7.14)$$

Proof. It follows from [32, Theorem 5] that for all t, x,

$$\nabla_\theta V(t, x; \pi_\theta) = \mathbb{E}\left[\int_t^T e^{\alpha(s-t)}\left\{\nabla_\theta \log \pi_\theta(a_s^{\pi_\theta}|s, X_s^\theta)\left(dV(s, X_s^\theta; \pi_\theta)\right.\right.\right.$$
$$\left.\left.\left. + \left[r_R(s, X_s^\theta, a_s^{\pi_\theta}) + \alpha V(s, X_s^\theta; \pi_\theta)\right]ds\right)\right\}\,\middle|\, X_t^\theta = x\right]. \quad (7.15)$$

By applying Itô's formula to $V(t, X_t^\theta; \pi_\theta)$, we have:

$$dV(t, X_t^\theta; \pi_\theta) = \left[\partial_t V(t, X_t^\theta; \pi_\theta) + \frac{1}{2}\sigma^2(t)\Delta V(t, X_t^\theta; \pi_\theta)\right.$$
$$\left. + b(t, X_t^\theta, a_t^{\pi_\theta}) \cdot \frac{\partial V}{\partial x}(t, X_t^\theta, \pi_\theta)\right]dt + \frac{\partial V}{\partial x}(t, X_t^\theta; \pi_\theta)\sigma(t)dB_t. \quad (7.16)$$

Injecting (7.16) into (7.15), and noting that $b(t, X_t^\theta, a_t^{\pi_\theta}) \cdot \frac{\partial V}{\partial x}(t, X_t^\theta, \pi_\theta) + r_R(t, X_t^\theta, a_t^{\pi_\theta})$ are the only terms involving $a_s^{\pi_\theta}$, we get

$$\nabla_\theta V(t, x; \pi_\theta) = \mathbb{E}\left[\int_t^T e^{\alpha(s-t)}\left\{\nabla_\theta \log \pi_\theta(a_s^{\pi_\theta}|s, X_s^\theta)\left(b(t, X_t^\theta, a_t^{\pi_\theta}) \cdot \frac{\partial V}{\partial x}(t, X_t^\theta, \pi_\theta)\right.\right.\right.$$
$$\left.\left.\left. + r_R(t, X_t^\theta, a_t^{\pi_\theta})ds\right)\right\}\,\middle|\, X_t^\theta = x\right].$$

Taking $t = 0$ yields the result. $\qquad\square$

Relying on Theorem 7.3, the machinery in [77] provides various algorithms (e.g., proximal policy optimization (PPO) [59]) to implement score matching.

8. Conclusion and future directions

In this section, we collect a few open problems which we hope to trigger future research.

(1) Most convergence results were poved for OU and VP. Empirical results show that VE (with $g(t) = \sqrt{2t}$) has the best performance to some dataset (e.g., CIFAR-10). It would be interesting to establish sharper results for VE, and explain why VE outperforms other models.

(2) In Section 5.2, VP is shown to have advantage if T is large, and the target data is strongly log-concave. In practice, one picks the time T, and tunes other model parameters carefully. This calls for considering (and comparing) various diffusion models under different scales.

(3) The theory for the ODE samplers, and especially the consistency model is sparse. [44] studied the convergence rate of a modified CD. It would be useful to establish provable/sharp guarantees for the original CD.

(4) There has also been work on diffusion models in constraint domains [16, 19, 46, 56]. Establishing the convergence result in each particular case is of independent interest.

(5) Approximation efficiency for score matching were studied in simple scenarios, e.g., exponential families, two-layer neural nets. It would be interesting to see whether one can go beyond these cases to match the practice.

(6) Diffusion guidance [15, 28] and fine-tuning [71, 72] are used to generate data that is customized for specific downstream tasks. Some recent work [24, 69, 75] explored the theoretical aspects of these approaches, which calls for further development.

(7) Diffusion model alignment is a promising subject in the context of RL with human feedback (RLHF). Rudiments are provided in Section 7.3, and it remains to build an entire framework, with provable guarantees.

References

[1] B. D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Process. Appl.*, 12(3):313–326, 1982.

[2] J. Benton, V. De Bortoli, A. Doucet, and G. Deligiannidis. Linear convergence bounds for diffusion models via stochastic localization. In *ICLR*, 2024.

[3] J. Benton, G. Deligiannidis, and A. Doucet. Error bounds for flow matching methods. 2023. arXiv:2305.16860.

[4] K. Black, M. Janner, Y. Du, I. Kostrikov, and S. Levine. Training diffusion models with reinforcement learning. In *ICLR*, 2024.

[5] A. Borji. Pros and cons of GAN evaluation measures. *Comput. Vis. Image Underst.*, 179:41–65, 2019.

[6] P. Cattiaux, G. Conforti, I. Gentil, and C. Léonard. Time reversal of diffusion processes under a finite entropy condition. *Ann. Inst. Henri Poincaré Probab. Stat.*, 59(4):1844–1881, 2023.

[7] M. Chen, K. Huang, T. Zhao, and M. Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *ICML*, volume 40, pages 4672–4712, 2023.

[8] M. Chen, S. Mei, J. Fan, and M. Wang. An overview of diffusion models: Applications, guided generation, statistical rates and optimization. 2024. arXiv:2404.07771.

[9] S. Chen, S. Chewi, H. Lee, Y. Li, J. Lu, and A. Salim. The probability flow ODE is provably fast. In *Neurips*, volume 36, 2023.

[10] 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. 2023.

[11] S. Chen, G. Daras, and A. Dimakis. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for ddim-type samplers. In *ICML*, volume 40, pages 4462–4484, 2023.

[12] A. S. Cherny and H.-J. Engelbert. *Singular stochastic differential equations*, volume 1858 of *Lecture Notes in Mathematics*. Springer-Verlag, 2005.

[13] V. De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. 2022. arXiv:2208.05314.

[14] 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.

[15] P. Dhariwal and A. Nichol. Diffusion models beat GANs on image synthesis. In *Neurips*, volume 34, pages 8780–8794, 2021.

[16] 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. 2022. arXiv preprint arXiv:2211.15089.

[17] B. Efron. Tweedie's formula and selection bias. *J. Amer. Statist. Assoc.*, 106(496):1602–1614, 2011.

[18] Y. Fan and K. Lee. Optimizing DDPM sampling with shortcut fine-tuning. 2023. arXiv:2301.13362.

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

[20] 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.

[21] 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.

[22] X. Gao, H. M. Nguyen, and L. Zhu. Wasserstein convergence guarantees for a general class of score-based generative models. 2023. arXiv:2311.11003.

[23] X. Gao and L. Zhu. Convergence analysis for general probability flow ODEs of diffusion models in Wasserstein distances. 2024. arXiv:2401.17958.

[24] Y. Guo, H. Yuan, Y. Yang, M. Chen, and M. Wang. Gradient guidance for diffusion models: An optimization perspective. 2024. arXiv:2404.14743.

[25] Y. Han, M. Razaviyayn, and R. Xu. Neural network-based score estimation in diffusion models: Optimization and generalization. In *ICLR*, 2024.

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

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

[28] J. Ho and T. Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[29] C.-W. Huang, J. H. Lim, and A. C. Courville. A variational perspective on diffusion-based generative models and score matching. In *Neurips*, volume 34, pages 22863–22876, 2021.

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

[31] A. Hyvärinen. Some extensions of score matching. *Comput. Statist. Data Anal.*, 51(5):2499–2512, 2007.

[32] Y. Jia and X. Y. Zhou. Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *J. Mach. Learn. Res.*, 23(154):1–55, 2022.

[33] Y. Jia and X. Y. Zhou. q-learning in continuous time. 2022. arXiv:2207.00713.

[34] 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.

[35] T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models. In *Neurips*, volume 35, pages 26565–26577, 2022.

[36] D. Kim, C.-H. Lai, W.-H. Liao, N. Murata, Y. Takida, T. Uesaka, Y. He, Y. Mitsufuji, and S. Ermon. Consistency trajectory models: Learning probability flow ODE trajectory of diffusion. In *ICLR*, 2024.

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

[38] F. Koehler, A. Heckett, and A. Risteski. Statistical efficiency of score matching: The view from isoperimetry. In *ICLR*, 2023.

[39] D. Kwon, Y. Fan, and K. Lee. Score-based generative modeling secretly minimizes the Wasserstein distance. In *Neurips*, volume 35, pages 20205–20217, 2022.

[40] H. Lee, J. Lu, and Y. Tan. Convergence for score-based generative modeling with polynomial complexity. In *Neurips*, volume 35, pages 22870–22882, 2022.

[41] H. Lee, J. Lu, and Y. Tan. Convergence of score-based generative modeling for general data distributions. In *ALT*, pages 946–985. PMLR, 2023.

[42] K. Lee, H. Liu, M. Ryu, O. Watkins, Y. Du, C. Boutilier, P. Abbeel, M. Ghavamzadeh, and S. S. Gu. Aligning text-to-image models using human feedback. 2023. arXiv:2302.12192.

[43] G. Li, Y. Huang, T. Efimov, Y. Wei, Y. Chi, and Y. Chen. Accelerating convergence of score-based diffusion models, provably. In *ICML*, 2024.

[44] G. Li, Z. Huang, and Y. Wei. Towards a mathematical theory for consistency training in diffusion models. 2024. arXiv:2402.07802.

[45] G. Li, Y. Wei, Y. Chen, and Y. Chi. Towards faster non-asymptotic convergence for diffusion-based generative models. In *ICLR*, 2024.

[46] A. Lou and S. Ermon. Reflected diffusion models. 2023. arXiv:2304.04740.

[47] C. Luo. Understanding diffusion models: A unified perspective. 2022. arXiv:2208.11970.

[48] S. Luo, Y. Tan, L. Huang, J. Li, and H. Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. 2023. arXiv:2310.04378.

[49] A. Millet, D. Nualart, and M. Sanz. Integration by parts and time reversal for diffusion processes. *Ann. Probab.*, 17(1):208–238, 1989.

[50] B. Ø ksendal. *Stochastic differential equations*. Universitext. Springer-Verlag, Berlin, fifth edition, 1998. An introduction with applications.

[51] OpenAI. Sora: Creating video from text. 2024. Available at `https://openai.com/sora`.

[52] C. Pabbaraju, D. Rohatgi, A. Sevekari, H. Lee, A. Moitra, and A. Risteski. Provable benefits of score matching. 2023. arXiv:2306.01993.

[53] J. Pidstrigach. Score-based generative models detect manifolds. In *Neurips*, volume 35, pages 35852–35865, 2022.

[54] 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.

[55] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. 2022. arXiv:2204.06125.

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

[57] H. Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I*, pages 157–163. Univ. California Press, Berkeley-Los Angeles, Calif., 1956.

[58] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[59] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. 2017. arXiv:1707.06347.

[60] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, volume 32, pages 2256–2265, 2015.

[61] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

[62] Y. Song and P. Dhariwal. Improved techniques for training consistency models. In *ICLR*, 2023.

[63] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever. Consistency models. In *ICML*, volume 40, page 32211–32252, 2023.

[64] Y. Song, C. Durkan, I. Murray, and S. Ermon. Maximum likelihood training of score-based diffusion models. In *Neurips*, volume 34, pages 1415–1428, 2021.

[65] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In *Neurips*, volume 32, page 11918–11930, 2019.

[66] Y. Song, S. Garg, J. Shi, and S. Ermon. Sliced score matching: A scalable approach to density and score estimation. In *UAI*, volume 35, pages 574–584, 2020.

[67] 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.

[68] D. W. Stroock and S. R. S. Varadhan. *Multidimensional diffusion processes*, volume 233 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, 1979.

[69] W. Tang. Fine-tuning of diffusion models via stochastic control: entropy regularization and beyond. 2024. arXiv:2403.06279.

[70] W. Tang and H. Zhao. Contractive diffusion probabilistic models. 2024. arXiv:2401.13115.

[71] M. Uehara, Y. Zhao, K. Black, E. Hajiramezanali, G. Scalia, N. L. Diamant, A. M. Tseng, T. Biancalani, and S. Levine. Fine-tuning of continuous-time diffusion models as entropy-regularized control. 2024. arXiv:2402.15194.

[72] M. Uehara, Y. Zhao, K. Black, E. Hajiramezanali, G. Scalia, N. L. Diamant, A. M. Tseng, S. Levine, and T. Biancalani. Feedback efficient online fine-tuning of diffusion models. 2024. arXiv:2402.16359.

[73] P. Vincent. A connection between score matching and denoising autoencoders. *Neural Comput.*, 23(7):1661–1674, 2011.

[74] H. Wang, T. Zariphopoulou, and X. Y. Zhou. Reinforcement learning in continuous time and space: A stochastic control approach. *J. Mach. Learn. Res.*, 21(198):1–34, 2020.

[75] Y. Wu, M. Chen, Z. Li, M. Wang, and Y. Wei. Theoretical insights for diffusion guidance: A case study for gaussian mixture models. 2024. arXiv:2403.01639.

[76] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Comput. Surv.*, 56(4):1–39, 2023.

[77] H. Zhao, W. Tang, and D. D. Yao. Policy optimization for continuous reinforcement learning. In *Neurips*, volume 36, 2023.

[78] H. Zhao, J. Zhang, D. D. Yao, and W. Tang. Score as action: tuning diffusion models by continuous reinforcement learning. 2024+. In preparation.

Department of Industrial Engineering and Operations Research, Columbia University.

Email address: `wt2319@columbia.edu`

Department of Industrial Engineering and Operations Research, Columbia University.

Email address: `hz2684@columbia.edu`