Journal of Machine Learning Research 26 (2025) 1-54 Submitted 6/24; Revised 1/25; Published 2/25

Wasserstein Convergence Guarantees for a General Class of Score-Based Generative Models

Xuefeng Gao XFGAO@SE.CUHK.EDU.HK
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
Shatin, NT, Hong Kong

Hoang M. Nguyen HMNGUYEN@FSU.EDU
Department of Mathematics
Florida State University
Tallahassee, FL 32306, United States of America

Lingjiong Zhu ZHU@MATH.FSU.EDU
Department of Mathematics
Florida State University
Tallahassee, FL 32306, United States of America

Editor: Jianfeng Lu

Abstract

Score-based generative models (SGMs) are a recent class of deep generative models with state-of-the-art performance in many applications. In this paper, we establish convergence guarantees for a general class of SGMs in the 2-Wasserstein distance, assuming accurate score estimates and smooth log-concave data distribution. We specialize our results to several concrete SGMs with specific choices of forward processes modeled by stochastic differential equations, and obtain an upper bound on the iteration complexity for each model, which demonstrates the impacts of different choices of the forward processes. We also provide a lower bound when the data distribution is Gaussian. Numerically, we experiment with SGMs with different forward processes for unconditional image generation on CIFAR-10. We find that the experimental results are in good agreement with our theoretical predictions on the iteration complexity.

Keywords: Score-based diffusion models, convergence analysis, Wasserstein distance, SDE-based sampler, iteration complexity

1. Introduction

Diffusion models are a powerful family of probabilistic generative models which can generate approximate samples from high-dimensional distributions (Sohl-Dickstein et al., 2015; Song and Ermon, 2019; Ho et al., 2020). The key idea in diffusion models is to use a forward process to progressively corrupt samples from a target data distribution with noise and then learn to reverse this process for generation of new samples. Diffusion models have achieved state-of-the-art performance in various applications such as image and audio generations, and they are the main components in popular content generators including Stable Diffusion (Rombach et al., 2022) and Dall-E 2 (Ramesh et al., 2022). We refer the readers to (Yang et al., 2023; Croitoru et al., 2023) for surveys on diffusion models.

A predominant formulation of diffusion models is score-based generative models (SGM) through Stochastic Differential Equations (SDEs) (Song et al., 2021), referred to as *Score SDEs* in Yang et al. (2023). At the core of this formulation, there are two stochastic processes in $\mathbb{R}^d$: a forward process and a reverse process. In this paper, we consider a general class of forward process $(\mathbf{x}_t)_{t\in[0,T]}$ described by the following SDE:

$$d\mathbf{x}_t = -f(t)\mathbf{x}_t dt + g(t)d\mathbf{B}_t, \quad \mathbf{x}_0 \sim p_0, \tag{1.1}$$

where both $f(t)$ and $g(t)$ are scalar-valued non-negative continuous functions of time t, $(\mathbf{B}_t)$ is the standard d-dimensional Brownian motion, and p_0 is the (unknown) target data distribution. The forward process has the interpretation of slowly injecting noise to data and transforming them to a noise-like distribution. If we reverse the forward process (1.1) in time, i.e., letting $(\tilde{\mathbf{x}}_t)_{t\in[0,T]} = (\mathbf{x}_{T-t})_{t\in[0,T]}$, then under mild assumptions, the reverse process $(\tilde{\mathbf{x}}_t)_{t\in[0,T]}$ satisfies another SDE (see e.g. Anderson (1982); Cattiaux et al. (2023)):

$$d\tilde{\mathbf{x}}_t = \left[f(T-t)\tilde{\mathbf{x}}_t + (g(T-t))^2\nabla_{\mathbf{x}}\log p_{T-t}(\tilde{\mathbf{x}}_t)\right]dt + g(T-t)d\bar{\mathbf{B}}_t, \quad \tilde{\mathbf{x}}_0 \sim p_T, \tag{1.2}$$

where $p_t(\mathbf{x})$ is the probability density function of $\mathbf{x}_t$ (the forward process at time t), $(\bar{\mathbf{B}}_t)$ is a standard Brownian motion in $\mathbb{R}^d$ and $\tilde{\mathbf{x}}_T = \mathbf{x}_0 \sim p_0$. Hence, the reverse process (1.2) transforms a noise-like distribution p_T into samples from p_0, which is the goal of generative modeling. Note that the reverse SDE (1.2) involves the *score function*, $\nabla_{\mathbf{x}}\log p_t(\mathbf{x})$, which is unknown. An important subroutine in SGM is to estimate this score function based on samples drawn from the forward process, typically by modeling time-dependent score functions as neural networks and training them on certain score matching objectives (Hyvärinen and Dayan, 2005; Vincent, 2011; Song et al., 2020). After the score is estimated, one can numerically solve the reverse SDE to generate new samples that approximately follows the data distribution (see Section 2). The Score-SDEs formulation is attractive because it generalizes and unifies several well-known diffusion models. In particular, the noise perturbation in Score Matching with Langevin dynamics (SMLD) (Song and Ermon, 2019) corresponds to the discretization of so-called variance exploding (VE) SDEs where $f \equiv 0$ in (1.1); The noise perturbation in Denoising Diffusion Probabilistic Modeling (DDPM) (Sohl-Dickstein et al., 2015; Ho et al., 2020) corresponds to an appropriate discretization of the variance preserving (VP) SDEs where $f(t) = \frac{1}{2}\beta(t)$ and $g(t) = \sqrt{\beta(t)}$ for some noise variance schedule $\beta(t)$, see Song et al. (2021) for details.

Despite the impressive empirical performances of diffusion models in various applications, the theoretical understandings of these models are relatively limited. In the past few years, there has been a rapidly growing body of literature on the convergence theory of diffusion models, assuming access to accurate estimates of the score function, see e.g. Block et al. (2020); De Bortoli (2022); Lee et al. (2022, 2023); Chen et al. (2023a,d); Benton et al. (2024); Chen et al. (2023b); Tang and Zhao (2024). While these studies have established polynomial convergence bounds for fairly general data distribution, this line of work has mostly focused on the analysis of the DDPM model where the corresponding forward SDE (1.1) satisfies $f = g^2/2 > 0$ (in fact many studies simply consider $f \equiv 1/2$ and $g \equiv 1$). However, it is important to understand the impacts of different choices of forward processes in diffusion models, which can potentially help provide theoretical guidance for selecting the functions f and g in the design of diffusion models. In this work, we make some progress towards addressing these issues using a theoretical approach based on convergence analysis. We summarize our contributions in the following.

Our Contributions.

- We establish convergence guarantees for a general class of SGMs in the 2-Wasserstein distance, assuming accurate score estimates and smooth log-concave data distribution (with unbounded support), see Theorem 2. In particular, we allow general functions f and g in the forward SDE (1.1). Theorem 2 directly translates to an upper bound on the iteration complexity, which is the number of sampling steps/iterations needed (in running the reverse process) to yield $\epsilon-$accuracy in the 2-Wasserstein distance between the data distribution and the generative distribution of the SGMs.

- We specialize our result to SGMs with specific functions f and g in the forward process. We find that under mild assumptions, the class of VP-SDEs (as forward processes) will lead to an iteration complexity bound $\widetilde{\mathcal{O}}\left(d/\epsilon^2\right)$ where $\widetilde{\mathcal{O}}$ ignores the logarithmic factors and hides dependency on other parameters. On the other hand, in the class of VE-SDEs, the choice of an exponential function g in Song et al. (2021) leads to an iteration complexity of $\widetilde{\mathcal{O}}\left(d/\epsilon^2\right)$, while other simple choices including polynomials for g lead to a worse complexity bound. We also find that VP-SDEs with polynomial and exponential noise schedules, which appear to be new to the literature, lead to better iteration complexity bounds (in terms of logarithmic dependence on d, ϵ) compared with the existing models. See Table 2 and Proposition 4.

- We also establish two new results on lower bounds. We first show that if we use the upper bound in Theorem 2, then in order to achieve ϵ accuracy, the iteration complexity is $\widetilde{\Omega}\left(d/\epsilon^2\right)$ for quite general functions f and g, where $\widetilde{\Omega}$ ignored the logarithmic dependence on ϵ and d (Proposition 5). This result, however, does not show whether our upper bound in Theorem 2 is tight or not. We next show that if the data distribution p_0 is Gaussian, then the lower bound for the iteration complexity is $\Omega\left(\sqrt{d}/\epsilon\right)$ (Proposition 6).

- Numerically, we experiment SGMs with different forward SDEs for unconditional image generation on the CIFAR-10 image dataset, using the neural network architectures

from Song et al. (2021). We find that the experimental results are in good agreement with our theoretical predictions: models with lower order of iteration complexity generally perform better, in the sense that they achieve lower FID scores and higher Inception scores (with the same stochastic sampler and number of sampling steps) over training iterations.

- Our main proof techniques (for the upper bound on the iteration complexity) rely on obtaining an explicit contraction rate for the reverse SDE in Wasserstein distance using properties of strongly log-concave distributions and Itô's formula, and controlling the discretization and score-matching errors by using synchronous coupling. This approach is significantly different from the existing studies on convergence analysis of SDE-based samplers such as Chen et al. (2023d), where Girsanov theorem and data processing inequality are used to obtain convergence guarantees in total variation distance or Kullback-Leibler (KL) divergence. Their techniques require the forward process to be contractive that excludes VE-SDEs, whereas our methodology enables us to cover a general class of models including both VP and VE-SDEs as special cases. We also emphasize that the reverse SDE is non-homogeneous in nature due to the score function, and one needs to perform a delicate analysis based on the result in Theorem 2 in order to spell out the leading order terms to obtain bounds on the iteration complexities for various VE-SDE and VP-SDE examples.

1.1 Related Work

The majority of the existing studies on the convergence analysis of diffusion models focus on the SDE-based implementation, that is, the sampling process for data generation is based on the discretization of the reverse-time SDE. For instance, (Lee et al., 2023; Chen et al., 2023a,d; Benton et al., 2024) have established polynomial convergence rates in Total Variation (TV) distance or KL divergence for the DDPM model, where they consider $f \equiv 1/2$ and $g \equiv 1$ in the forward SDE. In contrast to these studies, our work provides a unified convergence analysis for a more general class of diffusion models in the 2-Wasserstein distance ($\mathcal{W}_2$), thus illustrating how the choice of f and g impacts the iteration complexity. Focusing on $\mathcal{W}_2$ distance is not just of theoretical interest, but also of practice interest. Indeed, one of the most popular performance metrics for the quality of generated samples in image applications is Fréchet Inception Distance (FID), which measures the $\mathcal{W}_2$ distance between the distributions of generated images with the distribution of real images (Heusel et al., 2017). It is known that to obtain polynomial convergence rates in $\mathcal{W}_2$ distance, one often needs to assume some form of log-concavity for the data distribution, see e.g. (Chen et al., 2023d, Section 4) for discussions. Hence, we impose such (strong) assumptions on the data distribution. Although some studies (see e.g. Chen et al. (2023d)) also provided Wasserstein convergence bound for the DDPM model, they often assume that the data distribution is bounded, in which case the $\mathcal{W}_2$ distance can be bounded by the TV distance. In this work, we consider unbounded data distribution. We also emphasize that TV distance does not upper bound Wasserstein distance on $\mathbb{R}^d$ (see e.g. Gibbs and Su (2002)) and KL divergence does not imply Wasserstein convergence either (unless some additional conditions

are satisfied (Bolley and Villani, 2005)). Hence, our results are not implied by the existing convergence results for SDE-based samplers.

There is also a small but rapidly growing body of literature on convergence analysis of the probability flow ODE implementation of diffusion models (Song et al., 2021). See, e.g., Chen et al. (2023e,c); Li et al. (2024b); Gao and Zhu (2024); Li et al. (2024a). Our work differs from this line of studies in that we consider SDE-based samplers instead of ODE-based samplers.

Finally, our work is broadly related to the literature on the choice of noise schedules for diffusion models. In classical models, the choice of the forward process, i.e. f and g in (1.1), is often handcrafted and designed heuristically based on numerical performances of the corresponding models. For instance, the majority of existing DDPM models (in which $f = g^2/2$) use the linear noise schedule (i.e. $(g(t))^2 = b + at$ for some $a, b > 0$) proposed firstly in Ho et al. (2020). Nichol and Dhariwal (2021) proposed a cosine noise schedule and showed that it can improve the log-likelihood numerically. For the SMLD model, the noise schedule is often chosen as an exponential function, i.e. $g(t) = ab^t$ for some $a, b > 0$, following Song and Ermon (2019, 2020). In contrast to these studies, our work provides a theoretical analysis on the impact of different choices of forward processes based on the convergence analysis of diffusion models in $\mathcal{W}_2$ distance.

Notations. For any d-dimensional random vector $\mathbf{z}$ with finite second moment, the L_2-norm of $\mathbf{z}$ is defined as $\|\mathbf{z}\|_{L_2} = \left(\mathbb{E}\|\mathbf{z}\|^2\right)^{1/2}$, where $\|\cdot\|$ denotes the Euclidean norm. We denote $\mathcal{L}(\mathbf{z})$ as the law of $\mathbf{z}$. Define $\mathcal{P}_2(\mathbb{R}^d)$ as the space consisting of all the Borel probability measures ν on $\mathbb{R}^d$ with the finite second moment (based on the Euclidean norm). For any two Borel probability measures $\nu_1, \nu_2 \in \mathcal{P}_2(\mathbb{R}^d)$, the standard 2-Wasserstein distance Villani (2009) is defined by $\mathcal{W}_2(\nu_1, \nu_2) := \left(\inf \mathbb{E}\left[\|\mathbf{z}_1 - \mathbf{z}_2\|^2\right]\right)^{1/2}$, where the infimum is taken over all joint distributions of the random vectors $\mathbf{z}_1, \mathbf{z}_2$ with marginal distributions ν_1, ν_2. Finally, a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and L-smooth (i.e. ∇F is L-Lipschitz) if for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $\frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2 \geq F(\mathbf{x}) - F(\mathbf{y}) - \nabla F(\mathbf{y})^\top(\mathbf{x}-\mathbf{y}) \geq \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2$.

2. Preliminaries on SDE-Based Diffusion Models

We first recall the background on score-based generative modeling with SDEs (Song et al., 2021). Denote by $p_0 \in \mathcal{P}(\mathbb{R}^d)$ the unknown continuous data distribution, where $\mathcal{P}(\mathbb{R}^d)$ is the space of all probability measures on $\mathbb{R}^d$. Given i.i.d. samples from p_0, the goal is to generate new samples whose distribution closely resembles the data distribution.

Forward process and reverse process. Let $T > 0$. We consider a $d-$dimensional forward process $(\mathbf{x}_t)_{t\in[0,T]}$ given in (1.1). One can easily solve (1.1) to obtain

$$\mathbf{x}_t = e^{-\int_0^t f(s)ds}\mathbf{x}_0 + \int_0^t e^{-\int_s^t f(v)dv} g(s) d\mathbf{B}_s. \tag{2.1}$$

Denote by $p_t(\mathbf{x})$ the probability density of $\mathbf{x}_t$ for $t \geq 0$. Before we proceed, we give two popular classes of the forward SDEs in the literature (Song et al. (2021)):

- Variance Exploding (VE) SDE: $f(t) \equiv 0$ and $g(t) = \sqrt{\frac{d[\sigma^2(t)]}{dt}}$ for some nondecreasing function $\sigma(t)$, e.g., $g(t) = ae^{bt}$ for some positive constants a, b.
- Variance Preserving (VP) SDE: $f(t) = \frac{1}{2}\beta(t)$ and $g(t) = \sqrt{\beta(t)}$ for some nondecreasing function $\beta(t)$, e.g., $\beta(t) = at + b$ for some positive constants a, b.

Under mild assumptions, the reverse (in time) process $(\tilde{\mathbf{x}}_t)_{t\in[0,T]} = (\mathbf{x}_{T-t})_{t\in[0,T]}$ satisfies the SDE in (1.2). Hence, by starting from samples of p_T, we can run the SDE (1.2) to time T and obtain samples from the desired distribution p_0. However, the distribution p_T is not explicit and hard to sample from because of its dependency on the initial distribution p_0. With the choice of our forward SDE (1.1), which has an explicit solution (2.1) such that we can take

$$\hat{p}_T := \mathcal{N}\left(0, \int_0^T e^{-2\int_t^T f(s)ds}(g(t))^2 dt \cdot I_d\right), \tag{2.2}$$

as an approximation of p_T, where I_d is the $d-$dimensional identity matrix. Note that $\hat{p}_T$ is simply the distribution of the random variable $\int_0^t e^{-\int_s^t f(v)dv} g(s)d\mathbf{B}_s$ in (2.1), which is easy to sample from because it is Gaussian, and will be referred to as the prior distribution hereafter. Note that it directly follows from (2.1) that

$$\mathcal{W}_2(p_T, \hat{p}_T) \leq e^{-\int_0^T f(s)ds}\|\mathbf{x}_0\|_{L_2}. \tag{2.3}$$

In view of (1.2), we now consider the SDE:

$$d\mathbf{z}_t = \left[f(T-t)\mathbf{z}_t + (g(T-t))^2\nabla \log p_{T-t}(\mathbf{z}_t)\right]dt + g(T-t)d\bar{\mathbf{B}}_t, \quad \mathbf{z}_0 \sim \hat{p}_T. \tag{2.4}$$

Because $p_T \neq \hat{p}_T$, this creates an error that when running the reverse SDE, and as a result, the distribution of $\mathbf{z}_T$ differs from $\tilde{\mathbf{x}}_T \sim p_0$.

Remark. Several studies (see, e.g. De Bortoli (2022); Chen et al. (2023d)) consider VP-SDE, which is a time-inhomogeneous Ornstein-Uhlenbeck (OU) process, and use the stationary distribution p_∞ of the OU process as the prior distribution in their convergence analysis. If we use p_∞ instead of $\hat{p}_T$ in (2.2) as the prior distribution, our main result can be easily modified. Indeed, we only need to replace the error estimate (2.3) by an estimate on $\mathcal{W}_2(p_T, p_\infty)$, which can be easily bounded (by a coupling approach) due to the geometric convergence of the OU process to its stationarity. Because VE-SDE does not have a stationary distribution, we choose the prior distribution $\hat{p}_T$ in (2.2) in order to provide a unifying analysis based on the general forward SDE (1.1) that does not require the existence of a stationary distribution.

Score matching. Next, we consider score-matching. Note that the data distribution p_0 is unknown, and hence the true score function $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ in (2.4) is also unknown. In practice, it needs to be estimated/approximated by a time-dependent score model $s_\theta(\mathbf{x}, t)$,

which is often a deep neural network parameterized by θ. Estimating the score function from data has established methods, including score matching Hyvärinen and Dayan (2005), denoising score matching Vincent (2011), and sliced score matching Song et al. (2020). For instance, Song et al. (2021) use denoising score matching where the training objective for optimizing the neural network is given by

$$\min_{\theta} \mathbb{E}_{t\sim U[0,T]} \left[\lambda(t) \mathbb{E}_{\mathbf{x}_0} \mathbb{E}_{\mathbf{x}_t|\mathbf{x}_0} \left\| s_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p_{t|0}(\mathbf{x}_t|\mathbf{x}_0) \right\|^2 \right]. \tag{2.5}$$

Here, $\lambda(\cdot) : [0,T] \to \mathbb{R}_{>0}$ is some positive weighting function (e.g. $\lambda(t) = g(t)^2$), $U[0,T]$ is the uniform distribution on $[0,T]$, $\mathbf{x}_0 \sim p_0$ is the data distribution, and $p_{t|0}(\mathbf{x}_t|\mathbf{x}_0)$ is the density of $\mathbf{x}_t$ given $\mathbf{x}_0$. With the forward process in (1.1), we can easily infer from its solution (2.1) that the transition kernel $p_{t|0}(\mathbf{x}_t|\mathbf{x}_0)$ follows a Gaussian distribution where the mean the variance can be computed in closed form using f and g. Because we also have access to i.i.d. samples from p_0, the distribution of $\mathbf{x}_0$, the objective in (2.5) can be approximated by Monte Carlo methods in practice, and the resulting loss function can be then optimized. After the score function is estimated, we introduce a continuous-time process that approximates (2.4):

$$d\mathbf{u}_t = \left[f(T-t)\mathbf{u}_t + (g(T-t))^2 s_\theta(\mathbf{u}_t, T-t) \right] dt + g(T-t) d\bar{\mathbf{B}}_t, \quad \mathbf{u}_0 \sim \hat{p}_T, \tag{2.6}$$

where we replace the true score function in (2.4) by the estimated score s_θ.

Discretization and algorithm. To obtain an implementable algorithm, one can apply different numerical methods for solving the reverse SDE (2.6), see Section 4 of Song et al. (2021). In this paper, we consider the following Euler-type discretization of the continuous-time stochastic process (2.6). Let $\eta > 0$ be the stepsize and without loss of generality, let us assume that $T = K\eta$, where K is a positive integer. Let $\mathbf{y}_0 \sim \hat{p}_T$ and for any $k = 1, 2, \ldots, K$, we have

$$\begin{aligned} \mathbf{y}_k = \mathbf{y}_{k-1} &+ \left(\int_{(k-1)\eta}^{k\eta} f(T-t)dt \right) \mathbf{y}_{k-1} \\ &+ \left(\int_{(k-1)\eta}^{k\eta} (g(T-t))^2 dt \right) s_\theta(\mathbf{y}_{k-1}, T-(k-1)\eta) + \left(\int_{(k-1)\eta}^{k\eta} (g(T-t))^2 dt \right)^{1/2} \xi_k, \end{aligned} \tag{2.7}$$

where ξ_k are i.i.d. Gaussian random vectors $\mathcal{N}(0, I_d)$.

We are interested in the convergence of the generated distribution $\mathcal{L}(\mathbf{y}_K)$ to the data distribution p_0, where $\mathcal{L}(\mathbf{y}_K)$ denotes the law or distribution of $\mathbf{y}_K$. Specifically, our goal is to bound the 2-Wasserstein distance $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0)$, and investigate the number of iterates K that is needed in order to achieve ϵ accuracy, i.e. $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \le \epsilon$. At a high level, there are three sources of errors in analyzing the convergence: (1) the initialization of the algorithm at $\hat{p}_T$ instead of p_T, (2) the estimation error of the score function, and (3) the discretization error of the continuous-time process (2.6).

3. Main Results

In this section, we state our main results.

3.1 Assumptions

We first state our assumptions and provide some discussions regarding these assumptions.

Assumption 1 *Assume that p_0 is differentiable and positive everywhere. Moreover, $-\log p_0$ is m_0-strongly convex and L_0-smooth for some $m_0, L_0 > 0$.*

Remark 1 *Our strong-log-concave assumption on the (unbounded) data distribution has also been used in e.g. (Bruno et al., 2023; Gao and Zhu, 2024), which is a strong assumption. We need this assumption mainly because we consider Wasserstein convergence for score-based diffusions on a non-compact domain. In particular, we need the Wasserstein contractions in the reverse process to obtain convergence and control the discretization and score-matching errors, and this is achieved by deriving strong-log-concavity and smoothness for the score function at any time t based on this assumption. In the literature, some studies (Chen et al., 2023d) considered compactly supported data (without log-concavity) and establish Wasserstain convergence guarantees by early stopping the algorithm and projecting the algorithm output to the compact domain, whereas our analysis considers unbounded data distributions. It is an enormously interesting question how to relax Assumption 1. See the conclusion section for further discussions.*

Our next assumption is about the true score function $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ for $t \in [0, T]$. We assume that $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is Lipschitz in time, where the Lipschitz constant has at most linear growth in $\|\mathbf{x}\|$. Assumption 2 is needed in controlling the discretization error of the continuous-time process (2.6).

Assumption 2 *There exists some constant M_1 such that*

$$\sup_{1\le k\le K} \sup_{(k-1)\eta\le t\le k\eta} \left\|\nabla \log p_{T-t}(\mathbf{x}) - \nabla \log p_{T-(k-1)\eta}(\mathbf{x})\right\| \le M_1\eta(1+\|\mathbf{x}\|). \tag{3.1}$$

To motivate Assumption 2, consider the idealized case where $\mathbf{x}_0 \sim \mathcal{N}(0, \sigma_0^2 I_d)$. Then, one can compute that $\nabla_{\mathbf{x}} \log p_{T-t}(\mathbf{x}) = -\frac{\mathbf{x}}{(a_1(T-t))^2\sigma_0^2 + a_2(T-t)}$, where $a_1(T-t) := e^{-\int_0^{T-t} f(s)ds}$ and $a_2(T-t) := \int_0^{T-t} e^{-2\int_s^{T-t} f(v)dv}(g(s))^2 ds$. This implies that Assumption 2 is satisfied with $M_1 = \sup_{t\ge 0}\left|\frac{d}{dt}\frac{1}{(a_1(t))^2\sigma_0^2 + a_2(t)}\right| = \sup_{t\ge 0}\frac{|2a_1(t)a_1'(t)\sigma_0^2 + a_2'(t)|}{\left((a_1(t))^2\sigma_0^2 + a_2(t)\right)^2}$, provided that $M_1 \in (0, \infty)$. Indeed, for VE-SDE, $M_1 = \sup_{t\ge 0}\frac{(g(t))^2}{(\sigma_0^2 + \int_0^t (g(s))^2 ds)^2} < \infty$, provided that $(g(t))^2 \le c_1 +$

$c_2\left(\int_0^t (g(s))^2 ds\right)^2$ uniformly in t for some $c_1, c_2 > 0$, which is a very mild assumption that is satisfied by all our examples in Section 3.3. For VP-SDE, i.e. $f(t) = \frac{1}{2}\beta(t)$ and $g(t) = \sqrt{\beta(t)}$, then $a_2(t) = 1 - e^{-\int_0^t \beta(s)ds}$ and $M_1 = \sup_{t\geq 0} \frac{\beta(t)e^{-\int_0^t \beta(s)ds}|1-\sigma_0^2|}{\left(e^{-\int_0^t \beta(s)ds}(\sigma_0^2-1)+1\right)^2} < \infty$, provided that $\sup_{t\geq 0} \beta(t)e^{-\int_0^t \beta(s)ds} < \infty$ which is a very mild assumption that is satisfied by all our examples in Section 3.3.

We also make the following assumption on the score-matching approximation. Recall $(\mathbf{y}_k)$ are the iterates defined in (2.7).

Assumption 3 *Assume that there exists $M > 0$ such that*

$$\sup_{k=1,\ldots,K} \left\| \nabla \log p_{T-(k-1)\eta}\left(\mathbf{y}_{k-1}\right) - s_\theta\left(\mathbf{y}_{k-1}, T-(k-1)\eta\right)\right\|_{L_2} \leq M. \tag{3.2}$$

We make a remark here that the main results in our paper will still hold if Assumption 3 is to be replaced by the following L_2-type assumption on the score function of the continuous-time forward process $(\mathbf{x}_t)$ in (1.1):

$$\sup_{k=1,\ldots,K} \left\| \nabla \log p_{k\eta}\left(\mathbf{x}_{k\eta}\right) - s_\theta\left(\mathbf{x}_{k\eta}, k\eta\right)\right\|_{L_2} \leq M, \tag{3.3}$$

under the additional assumption that $s_\theta(\cdot, k\eta)$ is Lipschitz for every k and the observation that $\nabla \log p_{k\eta}(\cdot)$ is Lipschitz under Assumption 1 (see Lemma 9). Assumption (3.3) is considered in e.g. Chen et al. (2023d).

Finally, we make the following assumption on the stepsize η in the algorithm (2.7).

Assumption 4 *Assume that the stepsize η is small such that it satisfies the conditions:*

$$\eta \leq \min_{0\leq t\leq T} \left\{ \frac{\frac{(g(t))^2}{\frac{1}{m_0}e^{-2\int_0^t f(s)ds} + \int_0^t e^{-2\int_s^t f(v)dv}(g(s))^2 ds} - f(t)}{(f(t))^2 + (g(t))^4 (L(t))^2 + M_1 (g(t))^2} \right\}, \tag{3.4}$$

and

$$\eta \leq \min_{0\leq t\leq T} \left\{ \frac{1}{\frac{(g(t))^2}{\frac{1}{m_0}e^{-2\int_0^t f(s)ds} + \int_0^t e^{-2\int_s^t f(v)dv}(g(s))^2 ds} - f(t)} \right\}, \tag{3.5}$$

where for any $0 \leq t \leq T$,

$$L(t) := \min\left(\left(\int_0^t e^{-2\int_s^t f(v)dv} (g(s))^2 ds \right)^{-1}, \left(e^{\int_0^t f(s)ds} \right)^2 L_0 \right). \tag{3.6}$$

Note that in Assumption 4, $L(t)$ defined in (3.6) can be interpreted as the Lipschitz constant of $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ (Lemma 9). Under Assumption 4, the stepsize η is sufficiently small so that the discretization and score-matching errors will be controllable. For VE-SDE where $f = 0$, (3.4)-(3.5) can be simplified as $\eta \leq \min_{0\leq t\leq T} \frac{1}{(\frac{1}{m_0}+\int_0^t (g(s))^2 ds)((g(t))^2(L(t))^2+M_1)}$ and $\eta \leq \min_{0\leq t\leq T} \frac{\frac{1}{m_0}+\int_0^t (g(s))^2 ds}{(g(t))^2}$, where $L(t) = \min\left(\left(\int_0^t (g(s))^2 ds\right)^{-1}, L_0\right)$. On the other hand, For VP-SDEs where $f(t) = \frac{1}{2}\beta(t)$ and $g(t) = \sqrt{\beta(t)}$, (3.4)-(3.5) can be simplified as

$$\eta \leq \min_{0\leq t\leq T} \left\{ \frac{m_0 - (1-m_0)e^{-\int_0^t \beta(s)ds}}{(m_0 + (1-m_0)e^{-\int_0^t \beta(s)ds})(\frac{1}{2}\beta(t) + 2\beta(t)(L(t))^2 + 2M_1)} \right\}, \tag{3.7}$$

and

$$\eta \leq \min_{0\leq t\leq T} \frac{2}{\beta(t)} \cdot \frac{m_0 + (1-m_0)e^{-\int_0^t \beta(s)ds}}{m_0 - (1-m_0)e^{-\int_0^t \beta(s)ds}}, \tag{3.8}$$

where $L(t) = \min\left(\left(1 - e^{-\int_0^t \beta(s)ds}\right)^{-1}, e^{\int_0^t \beta(s)ds} L_0\right)$. To ensure that the right-hand sides of (3.4)-(3.5) are positive for general VP-SDEs, a sufficient condition is $m_0 > 1/2$.

3.2 Main Result

In this section, we state our main theoretical result, which provides a bound on the 2-Wasserstein distance $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0)$. To facilitate the presentation, we introduce a few quantities with their interpretations given in Table 1. For any $0 \leq t \leq T$, we define:

$$c(t) := \frac{m_0 (g(t))^2}{e^{-2\int_0^t f(s)ds} + m_0 \int_0^t e^{-2\int_s^t f(v)dv} (g(s))^2 ds}, \tag{3.9}$$

$$\begin{aligned}\mu(T-t) := {} & \frac{(g(T-t))^2}{\frac{1}{m_0} e^{-2\int_0^{T-t} f(s)ds} + \int_0^{T-t} e^{-2\int_s^{T-t} f(v)dv} (g(s))^2 ds} - f(T-t) \\ & - \eta(f(T-t))^2 - \eta(g(T-t))^4 (L(T-t))^2,\end{aligned} \tag{3.10}$$

$$m(t) := \frac{2(g(t))^2}{\frac{1}{m_0} e^{-2\int_0^t f(s)ds} + \int_0^t e^{-2\int_s^t f(v)dv} (g(s))^2 ds} - 2f(t), \tag{3.11}$$

$$c_1(T) := \sup_{0\leq t\leq T} e^{-\frac{1}{2}\int_0^t m(T-s)ds} e^{-\int_0^T f(s)ds} \|\mathbf{x}_0\|_{L_2}, \tag{3.12}$$

$$c_2(T) := \sup_{0\leq t\leq T} \left(e^{-2\int_0^t f(s)ds} \|\mathbf{x}_0\|_{L_2}^2 + d\int_0^t e^{-2\int_s^t f(v)dv} (g(s))^2 ds \right)^{1/2}, \tag{3.13}$$

and moreover for any $k = 1, 2, \ldots, K$,

$$\gamma_{k,\eta} := 1 - \int_{(k-1)\eta}^{k\eta} \mu(T-t)dt + M_1\eta \int_{(k-1)\eta}^{k\eta} (g(T-t))^2 dt, \tag{3.14}$$

$$h_{k,\eta} := c_1(T) \int_{(k-1)\eta}^{k\eta} [f(T-s) + (g(T-s))^2 L(T-s)]ds + c_2(T) \int_{T-k\eta}^{T-(k-1)\eta} f(s)ds + \left(\int_{T-k\eta}^{T-(k-1)\eta} (g(s))^2 ds \right)^{1/2} \sqrt{d}. \tag{3.15}$$

Quantities	Interpretations	Sources/References
$c(t)$ in (3.9)	Contraction rate of $\mathcal{W}_2(\mathcal{L}(\mathbf{z}_T), p_0)$	(5.1)
$L(T-t)$ in (3.6)	Lipschitz constant of $\nabla_{\mathbf{x}} \log p_{T-t}(\mathbf{x})$	Lemma 9
$\gamma_{k,\eta}$ in (3.14)	Contraction rate of discretization and score-matching errors in $\mathbf{y}_k$	Proposition 8
$m(t)$ in (3.11)	Contraction rate of $\mathbb{E}\|\tilde{\mathbf{x}}_T - \mathbf{z}_T\|^2$	(5.5)
$c_1(T)$ in (3.12)	Bound for $\sup_{0\le t\le T} \|\mathbf{z}_t - \tilde{\mathbf{x}}_t\|_{L_2}$	(A.6)
$c_2(T)$ in (3.13)	$\sup_{0\le t\le T} \|\mathbf{x}_t\|_{L_2}$	(5.14)
$h_{k,\eta}$ in (3.15)	Bound for $\sup_{(k-1)\eta\le t\le k\eta} \left\|\mathbf{z}_t - \mathbf{z}_{(k-1)\eta}\right\|_{L_2}$	Lemma 10

Table 1: Summary of quantities, their interpretations and the sources

We are now ready to state our bound on the 2-Wasserstein distance $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0)$.

Theorem 2 *Suppose that Assumptions 1, 2, 3 and 4 hold. Then, we have*

$$\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \le e^{-\int_0^{K\eta} c(t)dt} \|\mathbf{x}_0\|_{L_2} + \mathcal{E}_1(f, g, K, \eta, M_1) + \mathcal{E}_2(f, g, K, \eta, M, M_1), \tag{3.16}$$

where

$$\mathcal{E}_1(f, g, K, \eta, M_1) := \sum_{k=1}^{K} \prod_{j=k+1}^{K} \gamma_{j,\eta} \cdot \left(M_1\eta \left(1 + \|\mathbf{x}_0\|_{L_2} + c_2(T)\right) \int_{(k-1)\eta}^{k\eta} (g(T-t))^2 dt + \sqrt{\eta} h_{k,\eta} \left(\int_{(k-1)\eta}^{k\eta} [f(T-t) + (g(T-t))^2 L(T-t)]^2 dt \right)^{1/2} \right), \tag{3.17}$$

and

$$\mathcal{E}_2(f, g, K, \eta, M, M_1) := \sum_{k=1}^{K} \prod_{j=k+1}^{K} \gamma_{j,\eta} \cdot M \int_{(k-1)\eta}^{k\eta} (g(T-t))^2 dt, \tag{3.18}$$

where $c(t)$ is given in (3.9)*, $\gamma_{j,\eta}$ is defined in* (3.14)*, $c_2(T)$ is defined in* (3.13) *and $h_{k,\eta}$ is given in* (3.15)*.*

We can interpret Theorem 2 as follows. The first term in (3.16) is the *initialization error*; it characterizes the convergence of the continuous-time reverse SDE $(\mathbf{z}_t)$ in (2.4) to the distribution p_0 without discretization or score-matching errors. Specifically, it bounds the error $\mathcal{W}_2(\mathbf{z}_T, p_0)$, which is introduced due to the initialization of the reverse SDE $(\mathbf{z}_t)$ at $\hat{p}_T$ instead of p_T (see Proposition 7 for details). The second term $\mathcal{E}_1(f, g, K, \eta, M_1)$ and the third term $\mathcal{E}_2(f, g, K, \eta, M, M_1)$ in (3.16) quantify the *discretization error* and the *score-matching error* respectively in running the algorithm $(\mathbf{y}_k)$ in (2.7). Note that Assumption 4 implies that $\mu(T-t)$ in (3.10) is positive when η is sufficiently small, which further suggests from (3.14) that $\gamma_{j,\eta} \in (0, 1)$ for any $j = 1, 2, \ldots, K$. This quantity $\gamma_{j,\eta}$ that appears in the definitions of $\mathcal{E}_1(f, g, K, \eta, M_1)$ and $\mathcal{E}_2(f, g, K, \eta, M, M_1)$ is important, because it plays the role of a contraction rate of the error $\|\mathbf{z}_{k\eta} - \mathbf{y}_k\|_{L_2}$ over iterations (see Proposition 8 for details). Conceptually, it guarantees that as the number of iterations increases, the discretization and score-matching errors in the iterates $(\mathbf{y}_k)$ do not propagate and grow in time, which helps us control the overall discretization and score-matching errors.

3.3 Examples

In this section, we consider several examples of the forward processes and discuss the implications of Theorem 2. In particular, we consider a variety of choices for f and g in the SDE (1.1), and investigate the iteration complexity, i.e., the number of iterates K that is needed in order to achieve ϵ accuracy, i.e. $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq \epsilon$. While the bound in Theorem 2 is quite complex in general, and it can be made more explicit when we consider special f and g. We summarize the results in Table 2. The main idea behind the results in Table 2 is to analyze the three terms in (3.16) in Theorem 2 carefully for each example. We first choose $T = K\eta$ to be sufficiently large and fixed such that the first term in (3.16), that controls the initialization error, is $\mathcal{O}(\epsilon)$. Given $T = K\eta$ being fixed, the second term $\mathcal{E}_1$ in (3.16), that controls the discretization error, can be upper bounded by a function of $T = K\eta$ and η, which is $\mathcal{O}(\epsilon)$, by choosing η to be sufficiently small. This also determines K since $T = K\eta$ is chosen and fixed from the previous step. Finally, given $T = K\eta$ and η being fixed, the third term $\mathcal{E}_2$ in (3.16), that controls the score-matching error, can be upper bounded by a function of $T = K\eta$, η and M, which is $\mathcal{O}(\epsilon)$ by "choosing" M to be sufficiently small. For each example, with the specific choice of f and g, one needs to spell out the explicit dependence on $T = K\eta$, η and M from (3.16) before we can carry out the above analysis to obtain the results in Table 2. The detailed derivation of these results will be given in Appendix B.

From Table 2, we have the following observations. By ignoring the logarithmic dependence on ϵ, we can see that the VE-SDE from the literature Song and Ermon (2019) ($f(t) = 0$, $g(t) = ae^{bt}$ with $a, b > 0$), as well as all the VP-SDE examples in Table 2, achieve the complexity $\widetilde{\mathcal{O}}\left(d/\epsilon^2\right)$. In particular, the VP-SDEs that we proposed, i.e. $f(t) = \frac{1}{2}(b + at)^\rho$, $g(t) = (b + at)^{\rho/2}$ and $f(t) = \frac{1}{2}ae^{bt}$, $g(t) = \sqrt{a}e^{bt/2}$ have marginal improvement in terms of the logarithmic dependence on d and ϵ compared to the existing models in the literature. Among the VP-SDE models, $f(t) = \frac{1}{2}ae^{bt}$, $g(t) = \sqrt{a}e^{bt/2}$ has the best complexity performance. Indeed, the complexity for $f(t) = \frac{1}{2}(b + at)^\rho$, $g(t) = (b + at)^{\rho/2}$ is getting smaller

f	g	K	M	η	References
0	ae^{bt}	$\mathcal{O}\left(\frac{d\log(\frac{d}{\epsilon})}{\epsilon^2}\right)$	$\mathcal{O}\left(\frac{\epsilon}{\log(\frac{1}{\epsilon})}\right)$	$\mathcal{O}\left(\frac{\epsilon^2}{d}\right)$	Song et al. (2021)
0	a	$\mathcal{O}\left(\frac{d^{3/2}\log(\frac{d}{\epsilon})}{\epsilon^3}\right)$	$\mathcal{O}\left(\frac{\epsilon}{\sqrt{\log(\frac{d}{\epsilon})}}\right)$	$\mathcal{O}\left(\frac{\epsilon^2}{d\log(\frac{d}{\epsilon})}\right)$	De Bortoli et al. (2021)
0	$\sqrt{2at}$	$\mathcal{O}\left(\frac{d^{5/4}}{\epsilon^{5/2}}\right)$	$\mathcal{O}\left(\epsilon^{3/2}\right)$	$\mathcal{O}\left(\frac{\epsilon^2}{d}\right)$	our paper
0	$(b+at)^c$	$\mathcal{O}\left(\frac{d^{\frac{1}{2(2c+1)}+1}}{\epsilon^{\frac{1}{2c+1}+2}}\right)$	$\mathcal{O}\left(\epsilon^{1+\frac{2c}{2c+1}}\right)$	$\mathcal{O}\left(\frac{\epsilon^2}{d}\right)$	our paper
α	σ	$\mathcal{O}\left(\frac{d\log(\frac{d}{\epsilon})}{\epsilon^2}\right)$	$\mathcal{O}(\epsilon)$	$\mathcal{O}\left(\frac{\epsilon^2}{d}\right)$	De Bortoli et al. (2021)
$\frac{b+at}{2}$	$\sqrt{b+at}$	$\mathcal{O}\left(\frac{d\sqrt{\log(\frac{d}{\epsilon})}}{\epsilon^2}\right)$	$\mathcal{O}(\epsilon)$	$\mathcal{O}\left(\frac{\epsilon^2}{d}\right)$	Ho et al. (2020)
$\frac{(b+at)^\rho}{2}$	$(b+at)^{\frac{\rho}{2}}$	$\mathcal{O}\left(\frac{d(\log(\frac{d}{\epsilon}))^{\frac{1}{\rho+1}}}{\epsilon^2}\right)$	$\mathcal{O}(\epsilon)$	$\mathcal{O}\left(\frac{\epsilon^2}{d}\right)$	our paper
$\frac{ae^{bt}}{2}$	$\sqrt{a}e^{\frac{bt}{2}}$	$\mathcal{O}\left(\frac{d\log(\log(\frac{d}{\epsilon}))}{\epsilon^2}\right)$	$\mathcal{O}(\epsilon)$	$\mathcal{O}\left(\frac{\epsilon^2}{d}\right)$	our paper

Table 2: The iteration complexity of the algorithm (2.7) in terms of ϵ and dimension d. Here f, g correspond to the drift and diffusion terms in the forward SDE (1.1), and a, b, c, α, σ, ρ are positive constants. K is the number of iterates, M is the score-matching approximation error, and η is the stepsize required to achieve accuracy level ϵ (i.e. $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq \epsilon$).

as ρ increases. However, the complexity in Table 2 only highlights the dependence on d, ϵ and ignores the dependence on other model parameters including ρ. Therefore, in practice, we expect that for the performance of $f(t) = \frac{1}{2}(b+at)^\rho$, $g(t) = (b+at)^{\rho/2}$ is getting better when ρ is getting bigger, as long as it is below a certain threshold, since letting $\rho \to \infty$ will make the complexity K explode with its hidden dependence on ρ. This suggests that in practice, the optimal choice among the examples from Table 2 could be $f(t) = \frac{1}{2}(b+at)^\rho$, $g(t) = (b+at)^{\rho/2}$ for a reasonably large ρ. Later, in the numerical experiments (Section 4), we will see that this is indeed the case.

Another observation from Table 2 is that the complexity K has a phase transition, i.e. a discontinuity, when f decreases from α to 0 at $\alpha = 0$ (by considering the examples $f \equiv \alpha$, $g \equiv \sigma$ and $f \equiv 0$, $g \equiv a$), in the sense that the complexity jumps from $\mathcal{O}\left(\frac{d\log(d/\epsilon)}{\epsilon^2}\right)$ to $\mathcal{O}\left(\frac{d^{3/2}\log(d/\epsilon)}{\epsilon^3}\right)$. The intuition is that the drift term $\alpha > 0$ in the forward process creates a mean-reverting effect to make the starting point of the reverse process $\hat{p}_T$ closer to the idealized starting point p_T. Since we hide the dependence of complexity on α, and only keep track the dependence on ϵ, d, it creates a phase transition at $\alpha = 0$. A similar phase transition phenomenon occurs for the complexity K when we consider the examples $f \equiv 0$, $g(t) = ae^{bt}$ and $f \equiv 0$, $g(t) \equiv a$, where the complexity jumps from $\mathcal{O}\left(\frac{d\log(d/\epsilon)}{\epsilon^2}\right)$ to $\mathcal{O}\left(\frac{d^{3/2}\log(d/\epsilon)}{\epsilon^3}\right)$ as b decreases to 0.

Remark 3 *In the iteration complexities in Table 2, we omit the dependence on the constant M_1 from Assumption 2, where it is assumed to be a universal constant. One can easily include the dependence on M_1 for all the examples in Table 2, due to the explicit error bound we obtain in Theorem 2. For example, when $f(t) = 0$, $g(t) = ae^{bt}$, the iteration complexity K becomes $\mathcal{O}\left(\log\left(\frac{d}{\epsilon}\right)\max\left\{\frac{d}{\epsilon^2}, \frac{M_1\log(1/\epsilon)d^{3/4}}{\epsilon^{3/2}}\right\}\right)$, see Remark 13 in Appendix B; when $f(t) = \frac{b+at}{2}$, $g(t) = \sqrt{b+at}$, the iteration complexity K becomes $\mathcal{O}\left(\sqrt{\log\left(\frac{d}{\epsilon}\right)}\max\left\{\frac{d}{\epsilon^2}, \frac{M_1\sqrt{d}}{\epsilon}\right\}\right)$, see Remark 19 in Appendix B.*

3.4 Discussions

Our results in Table 2 naturally lead to several questions:

(1) For the VP-SDE examples in Table 2, we have seen that the iteration complexity is always of order $\widetilde{\mathcal{O}}\left(d/\epsilon^2\right)$ where $\widetilde{\mathcal{O}}$ ignores the logarithmic dependence on ϵ, d. One natural question is, for VP-SDE, is the complexity always of order $\widetilde{\mathcal{O}}\left(d/\epsilon^2\right)$?

(2) For our examples in Table 2, the best complexity is of order $\widetilde{\mathcal{O}}\left(d/\epsilon^2\right)$. Another natural question is that are there other choices of f, g so that the complexity becomes better than $\widetilde{\mathcal{O}}(d/\epsilon^2)$?

(3) If the answer to question (2) is negative, then are these upper bounds tight?

We have answers and results to these questions as follows.

First, we will show that the answer to question (1) is yes for a very wide class of general VP-SDE models. In particular, we show in in the next proposition that under mild assumptions on the function $\beta(t)$, the class of VP-SDEs (i.e., $f(t) = \frac{1}{2}\beta(t)$ and $g(t) = \sqrt{\beta(t)}$ for some nondecreasing function $\beta(t)$) will always lead to the complexity $\widetilde{\mathcal{O}}(\frac{d}{\epsilon^2})$ where $\widetilde{\mathcal{O}}$ ignores the logarithmic factors.

Proposition 4 *Under the assumptions of Theorem 2, assume that $\beta(t)$ is positive and increasing in t and there exist some $c_1, c_2, c_3 > 0$ such that $\beta(t) \le c_1\left(\int_0^t \beta(s)ds\right)^{c_3} + c_2 < \infty$ for every $t \ge 0$. Then, we have $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \le \mathcal{O}(\epsilon)$ after $K = \mathcal{O}\left(\frac{d(\log(d/\epsilon))^{3c_3+1}}{\epsilon^2}\right)$ iterations, provided that $M \le \frac{\epsilon}{(\log(\sqrt{d}/\epsilon))^{c_3}}$ and $\eta \le \frac{\epsilon^2}{d(\log(1/\epsilon))^{3/c_3}}$.*

It is easy to check that the assumptions in Proposition 4 are satisfied for all VP-SDEs examples in Table 2. Note that Proposition 4 is a general result for a wide class of VP-SDEs, and the dependence of the iteration complexity on the logarithmic factors of d and ϵ may be improved for various examples considered in Table 2. The details will be discussed and provided in Appendix B.2.

Next, we will show that the answer to question (2) is negative. In particular, we will show in the following proposition that if we use the upper bound (3.16) in Theorem 2, then in order to achieve ϵ accuracy, i.e. $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq \epsilon$, the complexity $K = \widetilde{\Omega}\left(d/\epsilon^2\right)$ under mild assumptions, where $\widetilde{\Omega}$ ignores the logarithmic dependence on ϵ and d.

Proposition 5 *Suppose the assumptions in Theorem 2 hold and we further assume that* $\min_{t\geq 0} g(t) > 0$ *and* $\min_{t\geq 0}(f(t)+(g(t))^2 L(t)) > 0$ *and* $\max_{0\leq s\leq t} c(s) \leq c_1 \left(\int_0^t c(s)ds\right)^{\rho} + c_2$ *uniformly in* t *for some* $c_1, c_2, \rho > 0$*, where* $c(s)$ *is defined in* (3.9)*. We also assume that* $\mu(t) \geq \frac{1}{4}m(t)$ *for any* $0 \leq t \leq T$ *(which holds for any sufficiently small* η*), where* $\mu(t), m(t)$ *are defined in* (3.10) *and* (3.11)*. If we use the upper bound* (3.16)*, then in order to achieve* ϵ *accuracy, i.e.* $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq \epsilon$*, we must have* $K = \widetilde{\Omega}\left(\frac{d}{\epsilon^2}\right)$*, where* $\widetilde{\Omega}$ *ignores the logarithmic dependence on* ϵ *and* d*.*

The assumptions in Proposition 5 are mild and one can readily check that they are satisfied for all the examples in Table 2. If we ignore the dependence on the logarithmic factors of d and ϵ, we can see from Table 2 that the VE-SDE example $f(t) \equiv 0$, $g(t) = ae^{bt}$, and all the VP-SDE examples achieve the lower bound in Proposition 5.

In Proposition 5, we showed that using the upper bound (3.16) in Theorem 2, we have the lower bound on the complexity $K = \widetilde{\Omega}\left(d/\epsilon^2\right)$. Therefore, the answer to question (2) is negative. This leads to question (3), which is, whether the iteration complexity $\tilde{\mathcal{O}}(d/\epsilon^2)$ (see Proposition 4 and Table 2) obtained from the upper bound in Theorem 2 is tight or not. This leads us to investigate a lower bound for the number of iterates that is needed to achieve ϵ accuracy. In the following proposition, we will show that the lower bound for the iteration complexity of algorithm (2.7) is at least $\Omega\left(\sqrt{d}/\epsilon\right)$ by constructing a special example when the initial distribution p_0 is Gaussian.

Proposition 6 *Consider the special case when* $\mathbf{x}_0$ *follows a Gaussian distribution* $\mathbf{x}_0 \sim \mathcal{N}(0, \sigma_0^2 I_d)$*. Then, in order to achieve the 2-Wasserstein* ϵ *accuracy, the iteration complexity has a lower bound* $\Omega\left(\frac{\sqrt{d}}{\epsilon}\right)$*, i.e. if there exists some* $T = T(\epsilon)$ *and* $\bar{\eta} = \bar{\eta}(\epsilon)$ *such that* $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq \epsilon$ *for any* $K \geq \bar{K} := T/\bar{\eta}$ *(with* $\eta = T/K \leq \bar{\eta}$*), then we must have* $\bar{K} = \Omega\left(\frac{\sqrt{d}}{\epsilon}\right)$*.*

The answer to question (3) is complicated. On the one hand, the lower bound $\Omega\left(\sqrt{d}/\epsilon\right)$ in Proposition 6 does not match the upper bound $\tilde{\mathcal{O}}(d/\epsilon^2)$. Note that the complexity $\Omega\left(\sqrt{d}/\epsilon\right)$ in Proposition 6 matches the upper bound for the complexity of an unadjusted Langevin algorithm under an additional assumption which is a growth condition on the third-order derivative of the log-density of the target distribution, see Li et al. (2022). However, under our current assumptions, it is shown in Dalalyan and Karagulyan (2019) that the upper bound for the complexity of an unadjusted Langevin algorithm matches the

upper bound $\tilde{\mathcal{O}}(d/\epsilon^2)$ in Table 2 that is deduced from Theorem 2. Hence, we speculate that the upper bound we obtained in Theorem 2 is tight under our current assumptions, and it may not be improvable unless additional assumptions are imposed. It will be left as a future research direction to explore whether under additional assumptions on the data distribution p_0, one can improve the upper bound in Theorem 2 and hence improve the complexity to match the lower bound $\Omega\left(\sqrt{d}/\epsilon\right)$ in Proposition 6, and furthermore, whether under the current assumptions, there exists an example other than the Gaussian distribution as illustrated in Proposition 6 that can match the upper bound $\tilde{\mathcal{O}}(d/\epsilon^2)$.

4. Numerical Experiments

In this section, we conduct numerical experiments based on various forward SDEs for unconditional image generation on the CIFAR-10 image dataset. Due to limitations in computational resources, the purpose of our experiments is not to beat or match the state-of-the-art numerical results such as FID scores. Instead, our goal is to compare the performances of diffusion models with different forward processes and better understand the impacts of such model choices numerically in addition to our theoretical findings.

4.1 SDEs for the Forward Process

We consider Variance Exploding (VE) SDEs and Variance Preserving (VP) SDEs as the forward processes in our experiments.

First, we consider various VE-SDEs which takes the form $d\mathbf{x}_t = \sqrt{\frac{d[\sigma^2(t)]}{dt}} d\mathbf{B}_t := g(t)d\mathbf{B}_t$, where $\sigma^2(t)$ is some non-decreasing function representing the scale of noise slowly added to the data over time $t \in [0,1]$. The transition kernel of VE-SDE is given by $p_{t|0}(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \mathbf{x}_0, \left[\sigma^2(t) - \sigma^2(0)\right] I_d\right)$. We consider several different noise functions $\sigma(t)$ (equivalently $g(t)$) below, and in the experiments we maintain the choice of $\sigma_{\min} := \sigma(0)$ and $\sigma_{\max} = \sigma(1)$ as in Song et al. (2021), where $\sigma_{\min} \ll \sigma_{\max}$.

(1) $g(t) = ab^t$ for some $a, b > 0$. (see Song and Ermon (2019); Song et al. (2021)). In this case, we have $\sigma(t) = \sigma_{\min}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^t$.

(2) $g(t) = \text{const}$, which yields $(\sigma(t))^2 = \sigma_{\min}^2 + (\sigma_{\max}^2 - \sigma_{\min}^2)t$.

(3) $g(t) = \sqrt{2at}$ for some constant $a > 0$, where $(\sigma(t))^2 = \sigma_{\min}^2 + (\sigma_{\max}^2 - \sigma_{\min}^2)t^2$.

(4) $g(t) = (b + at)^{\rho - \frac{1}{2}}$ for some $a, b > 0$, where $\sigma(t) = \left(\sigma_{\min}{}^{\frac{1}{\rho}} + \left(\sigma_{\max}{}^{\frac{1}{\rho}} - \sigma_{\min}{}^{\frac{1}{\rho}}\right) t\right)^{\rho}$. This noise schedule function $\sigma(t)$ is motivated by Karras et al. (2022), where they consider non-uniform discretization and the discretization/time steps are defined according to a sequence of noise levels based on $\sigma(t)$.

Next, we consider various VP-SDEs in the numerical experiments, where VP-SDE can be written as (see e.g. Song et al. (2021)) $d\mathbf{x}_t = -\frac{1}{2}\beta(t)\mathbf{x}_t dt + \sqrt{\beta(t)}d\mathbf{B}_t$, where $\beta(t)$ represents the noise scales over time $t \in [0,1]$. We consider the following choices of $\beta(t)$ in our experiments with $\beta(0) = \beta_{\min} \ll \beta(1) = \beta_{\max}$ (except the constant $\beta(t)$ case).

(1) $\beta(t) = b + at = \beta_{\min} + (\beta_{\max} - \beta_{\min})t$ for $t \in (0,1]$ (see Ho et al. (2020)).

(2) $\beta(t) = \beta_{\text{const}}$, which is a constant.

(3) $\beta(t) = (b+at)^\rho = \left(\beta_{\min}{}^{\frac{1}{\rho}} + \left(\beta_{\max}{}^{\frac{1}{\rho}} - \beta_{\min}{}^{\frac{1}{\rho}}\right)t\right)^\rho$ for $t \in (0,1]$.

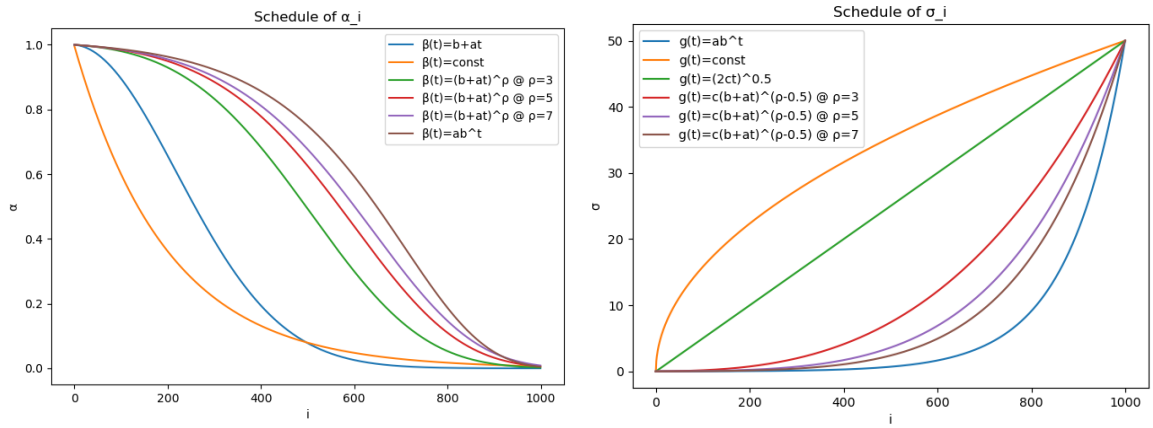
(4) $\beta(t) = ab^t = \beta_{\min} \cdot \left(\frac{\beta_{\max}}{\beta_{\min}}\right)^t$ for $t \in (0,1]$.

4.2 Experiment Setup

In this section we discuss the setup of the experiment.

Setup We focus on image generation with the "DDPM++ cont. (VP)" and "NCSN++ cont. (VE)" architectures from Song et al. (2021), whose generation processes correspond to the discretizations of the reverse-time VP-SDE and VE-SDE respectively. The models are trained on the popular 32×32 image dataset CIFAR-10, and we use the code base and structures in Song et al. (2021). However, we only have access to NVIDIA GeForce GTX 1080Ti, which has less available memory than the models in Song et al. (2021) requires; thus we reduce the number of channels in the residual blocks of DDPM++ and NCSN++ from 128 to 32, which effectively downsizes the filters dimension of the convolutional layers of the blocks by four times (see Appendix H in Song et al. (2021) for details of the neural network architecture). This will likely reduce the neural network's capability to capture more intricate details in the original data. However, since the purpose of our experiments is not to beat the state-of-the-art results, we focus on the performance comparison of different forward SDE models based on (non-deep) neural networks in Song et al. (2021) for score estimations. All models are trained for 3 million iterations (compared to 1.3 million iterations in Song et al. (2021)), since our models converge slower due to limitations of computing resources.

Relevant hyperparameters We choose the following configuration for the forward SDEs in the experiments. For VE-SDEs described in Section 4.1, we use $\sigma_{\min} = 0.01$ and $\sigma_{\max} = 50$ Song and Ermon (2020). For VP-SDEs, we choose $\beta_{\text{const}} = 0.005$ for constant $\beta(t)$; $\beta_{\min} = 10^{-4}$, $\beta_{\max} = 0.03$ for $\beta(t) = ab^t$, and $\beta_{\min} = 10^{-4}$, $\beta_{\max} = 0.02$ (Ho et al. (2020)) for the other VP-SDEs to maintain the progression of $\alpha_i := \prod_{j=1}^{i}(1-\beta_i)$, where $\{\beta_i\}_{i=1}^N$ is the discretization of $\beta(t)$. Note that α_i slowly progresses from $1-\beta_{\min}$ to 0 for all our VP-SDEs, and one can equivalently use α_i's to demonstrate the noise schedules instead of β_i (similar to Nichol and Dhariwal (2021)). Figure 1 shows the difference of α_i for DDPM models (corresponding to discretization of VP-SDEs) and σ_i (discretized noise level $\sigma(t)$) for NCSN (a.k.a. SMLD) models (corresponding to discretization of VE-SDEs).



(a) DDPM (b) NCSN

Figure 1: The schedules of (α_i) for DDPM models and (σ_i) for NCSN (a.k.a. SMLD) models with different forward SDEs.

Sampling We generate samples using the Euler-Maruyama solver for discretizing the reverse-time SDEs, referred to as the predictor in Song et al. (2021). We set the number of discretized time steps $N = 1000$, which follows Ho et al. (2020) and Song et al. (2021).

Performance Metrics The quality of the generated image samples is evaluated with Fréchet Inception Distance (FID, lower is better), which was first introduced by Heusel et al. (2017) for measuring the 2-Wasserstein distance between the distribution of generated images and the distribution of real images. We also report the Inception Score (IS) (see Salimans et al. (2016)) of the generated images as a secondary measure. However, IS (higher is better) only evaluates how realistic the generated images are without a comparison to real images. Each FID and IS measure is evaluated based on $20,000$ samples.

4.3 Empirical Results

Table 3 and Figures 2–3 show the performances of various diffusion models corresponding to different forward SDEs that we used. We have the following two important observations.

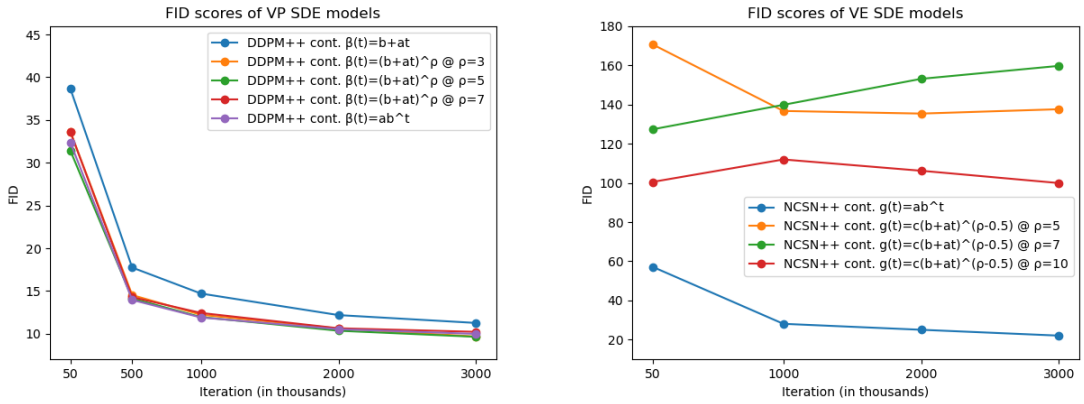
First, the experimental results are in good agreement with our theoretical prediction on the iteration complexity in Table 2. With the same number of discretized time steps, models (forward SDEs) with lower order of iteration complexity generally obtain a better FID score and IS (lower FID and higher IS) over training iterations. In addition, as predicted by the theory in Table 2, VE-SDE models generally perform worse than VP-SDEs. Among the VE-SDE models, the choice of $f \equiv 0$ and $g(t) = ab^t$ leads to the best performance in terms of FID and IS scores. We also remark that VE-SDE models can perform significantly better with a corrector (see Song et al. (2021) for Predictor-Corrector sampling), and get close to the performance of VP-SDE models. However, our experimental results are based on the stochastic sampler without any corrector in order to fit the setup of our theory.

Second, our experimental results show that our proposed VP-SDE with a polynomial variance schedule $\beta(t) = (b+at)^\rho$ for some ρ or an exponential variance schedule $\beta(t) = ab^t$ can outperform the other existing models, at least with simpler neural network architectures. The optimal ρ is around 5 according to Table 3. This is again consistent with our discussion in Section 3.3.

We also choose the best performing models from Table 3 and test the deeper neural network architecture in Song et al. (2021) (which doubles the number of residual blocks per resolution, except for reduced batch size and reduced number of filters due to our memory limitation, as mentioned in Section 4.2). The results are shown in Figure 4 and Table 4. We can see that the performance of different models remains consistent in a more complex architecture setting.

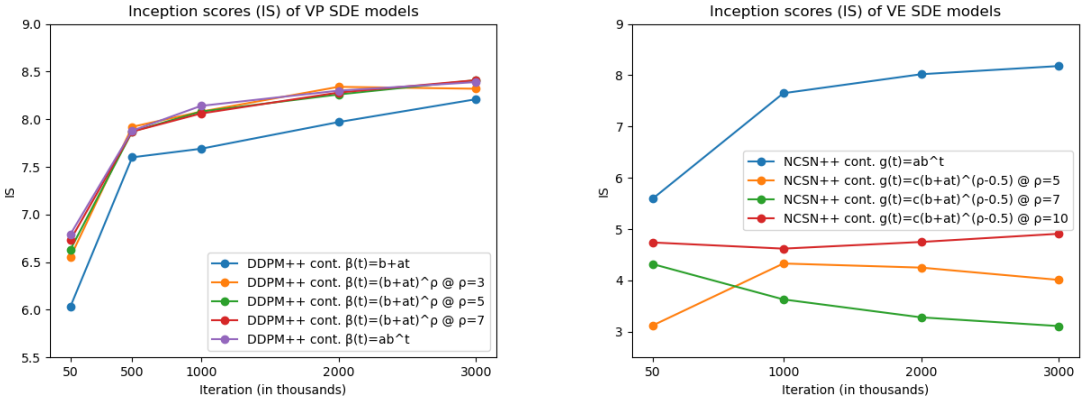
Model	**FID**↓	**IS**↑	References
DDPM (VP - $\beta(t) = \text{const}$)	17.46	8.19	De Bortoli et al. (2021)
DDPM (VP - $\beta(t) = b + at$)	11.26	8.21	Ho et al. (2020)
DDPM (VP - $\beta(t) = (b + at)^\rho, \rho = 2$)	9.77	8.33	our paper
DDPM (VP - $\beta(t) = (b + at)^\rho, \rho = 3$)	9.67	8.32	
DDPM (VP - $\beta(t) = (b + at)^\rho, \rho = 5$)	9.64	8.41	
DDPM (VP - $\beta(t) = (b + at)^\rho, \rho = 7$)	10.22	8.41	
DDPM (VP - $\beta(t) = (b + at)^\rho, \rho = 10$)	10.27	8.51	
DDPM (VP - $\beta(t) = ab^t$)	9.98	8.39	our paper
NCSN (VE - $g(t) = ab^t$)	22.11	8.18	Song et al. (2021)
NCSN (VE - $g(t) = \text{const}$)	461.42	1.18	De Bortoli et al. (2021)
NCSN (VE - $g(t) = \sqrt{2at}$)	457.04	1.20	our paper
NCSN (VE - $g(t) = (b + at)^{\rho-\frac{1}{2}}, \rho = 2$)	369.51	1.34	our paper
NCSN (VE - $g(t) = (b + at)^{\rho-\frac{1}{2}}, \rho = 3$)	233.20	1.95	
NCSN (VE - $g(t) = (b + at)^{\rho-\frac{1}{2}}, \rho = 5$)	137.55	4.01	
NCSN (VE - $g(t) = (b + at)^{\rho-\frac{1}{2}}, \rho = 7$)	159.66	3.11	
NCSN (VE - $g(t) = (b + at)^{\rho-\frac{1}{2}}, \rho = 10$)	99.89	4.91	

Table 3: Performances of different SDE models on CIFAR-10 at 3,000,000 iterations



(a) DDPM (b) NCSN (aka SMLD)

Figure 2: The FID score progressions of different SDE models on CIFAR-10



(a) DDPM (b) NCSN (aka SMLD)

Figure 3: The inception score (IS) progressions of different SDE models on CIFAR-10

5. Analysis: Proofs of the Main Results

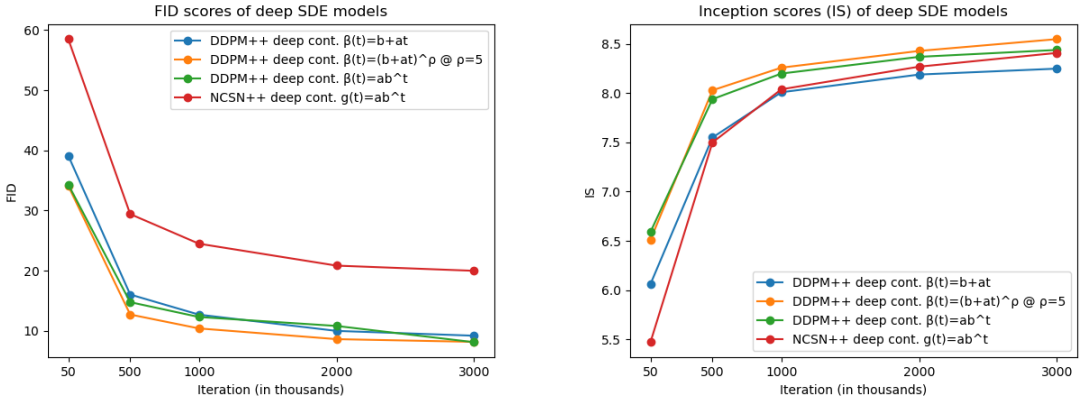
5.1 Proof of Theorem 2

To prove Theorem 2, we study the three sources of errors discussed in Section 2: (1) the initialization of the algorithm at $\hat{p}_T$ instead of p_T, (2) the estimation error of the score function, and (3) the discretization error of the continuous-time process (2.6).

First, we study the error introduced due to the initialization at $\hat{p}_T$ instead of p_T. Recall the reverse SDE $\mathbf{z}_t$ given in (2.4). As discussed in Section 2, the distribution of $\mathbf{z}_T$ differs from p_0, because $\mathbf{z}_0 \sim \hat{p}_T \neq p_T$. The following result provides a bound on $\mathcal{W}_2(\mathcal{L}(\mathbf{z}_T), p_0)$.

Model	FID↓	IS↑
DDPM deep (VP - $\beta(t) = b + at$)	9.22	8.25
DDPM deep (VP - $\beta(t) = (b + at)^\rho, \rho = 5$)	8.20	8.55
DDPM deep (VP - $\beta(t) = ab^t$)	8.14	8.44
NCSN deep (VE - $g(t) = ab^t$)	20.00	8.41

Table 4: Performances of deep versions of the best performing models from Table 3 on CIFAR-10 at 3,000,000 iterations. SMLD is also known as NCSN.



(a) FID (b) IS

Figure 4: The FID and IS score progressions of the deep version of the best performing SDE models on CIFAR-10

Proposition 7 *Assume that p_0 is m_0-strongly-log-concave. Then, we have*

$$\mathcal{W}_2(\mathcal{L}(\mathbf{z}_T), p_0) \leq e^{-\int_0^T c(t)dt} \|\mathbf{x}_0\|_{L_2}, \tag{5.1}$$

where $c(t)$ is given in (3.9).

The main challenge in analyzing the SDE $\mathbf{z}_t$ lies in studying the term $\nabla \log p_{T-t}(\mathbf{z}_t)$. In general, this term is neither linear in $\mathbf{z}_t$ nor admits a closed-form expression. However, when p_0 is strongly log-concave, we are able to show that $\log p_{T-t}(\mathbf{x})$ is also strongly concave. This fact, together with Itô's formula for SDEs, allows us to establish Proposition 7. The proof of Proposition 7 is given in Section 5.1.2.

Now we consider the algorithm (2.7) with iterates $(\mathbf{y}_k)$, and bound the errors due to score estimations and discretizations together. For any $k = 0, 1, 2, \ldots, K$, $\mathbf{y}_k$ has the same distribution as $\hat{\mathbf{y}}_{k\eta}$, where $\hat{\mathbf{y}}_t$ is a continuous-time process with the dynamics:

$$d\hat{\mathbf{y}}_t = \left[f(T-t)\hat{\mathbf{y}}_{\lfloor t/\eta \rfloor \eta} + (g(T-t))^2 s_\theta\left(\hat{\mathbf{y}}_{\lfloor t/\eta \rfloor \eta}, T - \lfloor t/\eta \rfloor \eta\right)\right] dt + g(T-t)d\bar{\mathbf{B}}_t, \tag{5.2}$$

with $\hat{\mathbf{y}}_0 \sim \hat{p}_T$. We have the following result that provides an upper bound for $\|\mathbf{z}_{k\eta} - \hat{\mathbf{y}}_{k\eta}\|_{L_2}$ in terms of $\|\mathbf{z}_{(k-1)\eta} - \hat{\mathbf{y}}_{(k-1)\eta}\|_{L_2}$. This result plays a key role in the proof of Theorem 2.

Proposition 8 *Assume that p_0 is m_0-strongly-log-concave, i.e. $-\log p_0$ is m_0-strongly convex and $\nabla \log p_0$ is L_0-Lipschitz. For any $k = 1, 2, \ldots, K$,*

$$\begin{aligned}\|\mathbf{z}_{k\eta} - \hat{\mathbf{y}}_{k\eta}\|_{L_2} &\le \gamma_{k,\eta}\|\mathbf{z}_{(k-1)\eta} - \hat{\mathbf{y}}_{(k-1)\eta}\|_{L_2} \\ &\quad + M_1\eta\left(1 + \|\mathbf{x}_0\|_{L_2} + c_2(T)\right)\int_{(k-1)\eta}^{k\eta}(g(T-t))^2dt + M\int_{(k-1)\eta}^{k\eta}(g(T-t))^2dt \\ &\quad + \sqrt{\eta}h_{k,\eta}\left(\int_{(k-1)\eta}^{k\eta}[f(T-t) + (g(T-t))^2L(T-t)]^2dt\right)^{1/2}, \end{aligned} \tag{5.3}$$

where $\gamma_{k,\eta}$ is defined in (3.14)*, $c_2(T)$ is defined in* (3.13) *and $h_{k,\eta}$ is given in* (3.15)*.*

We remark that the coefficient $\gamma_{j,\eta}$ in front of the term $\left\|\mathbf{z}_{(k-1)\eta} - \hat{\mathbf{y}}_{(k-1)\eta}\right\|_{L_2}$ in (5.3) lies in between zero and one. Indeed, it follows from Assumption 4 and the definition of $\mu(t)$ in (3.10) that $\mu(t) \ge M_1\eta(g(t))^2$ for every $0 \le t \le T$ and $\eta\max_{0\le t\le T}\mu(t) < 1$ such that for any $j = 1, 2, \ldots, K$, $0 \le \gamma_{j,\eta} \le 1$ where $\gamma_{j,\eta}$ is defined in (3.14).

Now we are ready to prove Theorem 2.

5.1.1 Completing the Proof of Theorem 2

Proof Since $\hat{\mathbf{y}}_{k\eta}$ has the same distribution as $\mathbf{y}_k$, by applying (5.3), we have

$$\begin{aligned}&\mathcal{W}_2(\mathcal{L}(\mathbf{z}_{K\eta}), \mathcal{L}(\mathbf{y}_K)) \le \|\mathbf{z}_{K\eta} - \hat{\mathbf{y}}_{K\eta}\|_{L_2} \\ &\le \sum_{k=1}^{K}\prod_{j=k+1}^{K}\gamma_{j,\eta}\cdot\Bigg(M_1\eta\left(1 + \|\mathbf{x}_0\|_{L_2} + c_2(T)\right)\int_{(k-1)\eta}^{k\eta}(g(T-t))^2dt \\ &\qquad\qquad + M\int_{(k-1)\eta}^{k\eta}(g(T-t))^2dt \\ &\qquad + \sqrt{\eta}h_{k,\eta}\left(\int_{(k-1)\eta}^{k\eta}\left[f(T-t) + (g(T-t))^2L(T-t)\right]^2dt\right)^{1/2}\Bigg).\end{aligned}$$

Moreover, we recall that $T = K\eta$ and by triangle inequality for 2-Wasserstein distance, $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \le \mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), \mathcal{L}(\mathbf{z}_{K\eta})) + \mathcal{W}_2(\mathcal{L}(\mathbf{z}_{K\eta}), p_0)$. The proof is completed by applying Proposition 7. ■

5.1.2 Proof of Proposition 7

Proof For the forward SDE (1.1), the transition density is Gaussian, and we have $p_t(\mathbf{x}_t) = \int_{\mathbb{R}^d} p(\mathbf{x}_t|\mathbf{x}_0)p_0(\mathbf{x}_0)d\mathbf{x}_0$, where

$$p(\mathbf{x}_t|\mathbf{x}_0) = \frac{1}{\left(2\pi \int_0^t e^{-2\int_s^t f(v)dv}(g(s))^2 ds\right)^{d/2}} \exp\left(-\frac{\|\mathbf{x}_t - e^{-\int_0^t f(s)ds}\mathbf{x}_0\|^2}{2\int_0^t e^{-2\int_s^t f(v)dv}(g(s))^2 ds}\right).$$

This implies that

$$\begin{aligned}
\log p_{T-t}(\mathbf{x}) &= \log \int_{\mathbb{R}^d} \exp\left(-\frac{\|\mathbf{x} - e^{-\int_0^{T-t} f(s)ds}\mathbf{x}_0\|^2}{2\int_0^{T-t} e^{-2\int_s^{T-t} f(v)dv}(g(s))^2 ds}\right) p_0(\mathbf{x}_0)d\mathbf{x}_0 \\
&\qquad - \frac{d}{2}\log\left(2\pi \int_0^t e^{-2\int_s^t f(v)dv}(g(s))^2 ds\right) \\
&= \log \int_{\mathbb{R}^d} \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}_0\|^2}{2\int_0^{T-t} e^{-2\int_s^{T-t} f(v)dv}(g(s))^2 ds}\right) p_0\left(e^{\int_0^{T-t} f(s)ds}\mathbf{x}_0\right) d\mathbf{x}_0 \\
&\qquad + de^{\int_0^{T-t} f(s)ds} - \frac{d}{2}\log\left(2\pi \int_0^t e^{-2\int_s^t f(v)dv}(g(s))^2 ds\right),
\end{aligned}$$

where we applied change-of-variable to obtain the last equation. Note that for any two functions $p, q : \mathbb{R}^d \to \mathbb{R}$, where p is m_p-strongly-log-concave and q is m_q-strongly-log-concave (i.e. $-\log p$ is m_p-strongly-convex and $-\log q$ is m_q-strongly-convex) then it is known that the convolution of p and q, i.e. $\int_{\mathbb{R}^d} p(\mathbf{x}-\mathbf{y})q(\mathbf{y})d\mathbf{y}$ is $(m_p^{-1} + m_q^{-1})^{-1}$-strongly-log-concave; see e.g. Proposition 7.1 in Saumard and Wellner (2014). It is easy to see that the function $\mathbf{x} \mapsto \exp\left(-\frac{\|\mathbf{x}\|^2}{2\int_0^{T-t} e^{-2\int_s^{T-t} f(v)dv}(g(s))^2 ds}\right)$ is $\frac{1}{\int_0^{T-t} e^{-2\int_s^{T-t} f(v)dv}(g(s))^2 ds}$-strongly-log-concave, and the function $\mathbf{x} \mapsto p_0\left(e^{\int_0^{T-t} f(s)ds}\mathbf{x}\right)$ is $m_0\left(e^{\int_0^{T-t} f(s)ds}\right)^2$-strongly-log-concave since we assumed that $\mathbf{x} \mapsto p_0(\mathbf{x})$ is m_0-strongly-log-concave. Hence, we conclude that $\log p_{T-t}(\mathbf{x})$ is $a(T-t)$-strongly-concave, where

$$a(T-t) := \frac{1}{\frac{1}{m_0}e^{-2\int_0^{T-t} f(s)ds} + \int_0^{T-t} e^{-2\int_s^{T-t} f(v)dv}(g(s))^2 ds}. \tag{5.4}$$

Next, let us recall the definition of $m(T-t)$ in (3.11) and the dynamics of $\tilde{\mathbf{x}}_t$ and $\mathbf{z}_t$ in (1.2) and (2.4) respectively. By Itô's formula,

$$\begin{aligned}
&d\left(\|\tilde{\mathbf{x}}_t - \mathbf{z}_t\|^2 e^{\int_0^t m(T-s)ds}\right) \\
&= m(T-t)e^{\int_0^t m(T-s)ds}\|\tilde{\mathbf{x}}_t - \mathbf{z}_t\|^2 dt + 2e^{\int_0^t m(T-s)ds}\langle \tilde{\mathbf{x}}_t - \mathbf{z}_t, d\tilde{\mathbf{x}}_t - d\mathbf{z}_t\rangle \\
&= m(T-t)e^{\int_0^t m(T-s)ds}\|\tilde{\mathbf{x}}_t - \mathbf{z}_t\|^2 dt + 2e^{\int_0^t m(T-s)ds}\langle \tilde{\mathbf{x}}_t - \mathbf{z}_t, f(T-t)(\tilde{\mathbf{x}}_t - \mathbf{z}_t)\rangle dt \\
&\quad + 2e^{\int_0^t m(T-s)ds}\left\langle \tilde{\mathbf{x}}_t - \mathbf{z}_t, (g(T-t))^2\left(\nabla \log p_{T-t}(\tilde{\mathbf{x}}_t) - \nabla \log p_{T-t}(\mathbf{z}_t)\right)\right\rangle dt \\
&\le e^{\int_0^t m(T-s)ds}\left(m(T-t) + 2f(T-t) - 2(g(T-t))^2 a(T-t)\right)\|\tilde{\mathbf{x}}_t - \mathbf{z}_t\|^2 dt = 0.
\end{aligned}$$

This implies that $\|\tilde{\mathbf{x}}_t - \mathbf{z}_t\|^2 e^{\int_0^t m(T-s)ds} \leq \|\tilde{\mathbf{x}}_0 - \mathbf{z}_0\|^2$, so that

$$\mathbb{E}\|\tilde{\mathbf{x}}_T - \mathbf{z}_T\|^2 \leq e^{-\int_0^T m(T-s)ds} \mathbb{E}\|\tilde{\mathbf{x}}_0 - \mathbf{z}_0\|^2. \tag{5.5}$$

Consider a coupling of $(\tilde{\mathbf{x}}_0, \mathbf{z}_0)$ such that $\tilde{\mathbf{x}}_0 \sim p_T$, $\mathbf{z}_0 \sim \hat{p}_T$ and $\mathbb{E}\|\tilde{\mathbf{x}}_0 - \mathbf{z}_0\|^2 = \mathcal{W}_2^2(p_T, \hat{p}_T)$. Together with (2.3), we conclude that

$$\begin{aligned}
\mathcal{W}_2^2(\mathcal{L}(\mathbf{z}_T), p_0) = \mathcal{W}_2^2(\mathcal{L}(\mathbf{z}_T), \mathcal{L}(\tilde{\mathbf{x}}_T)) &\leq \mathbb{E}\|\tilde{\mathbf{x}}_T - \mathbf{z}_T\|^2 \\
&\leq e^{-\int_0^T m(T-s)ds} \mathcal{W}_2^2(p_T, \hat{p}_T) \\
&\leq e^{-\int_0^T m(s)ds} e^{-2\int_0^T f(s)ds} \|\mathbf{x}_0\|_{L_2}^2 \\
&= e^{-2\int_0^T c(t)dt} \|\mathbf{x}_0\|_{L_2}^2.
\end{aligned}$$

The proof is complete. ■

5.1.3 Proof of Proposition 8

We first state a key technical lemma, which will be used in the proof of Proposition 8. The proof of the following result will be provided in Appendix A.1.

Lemma 9 *Suppose that Assumption 1 holds. Then, $\nabla_{\mathbf{x}} \log p_{T-t}(\mathbf{x})$ is $L(T-t)$-Lipschitz in $\mathbf{x}$, where $L(T-t)$ is given in (3.6).*

Proof By recalling the dynamics of $\mathbf{z}_t$ and $\hat{\mathbf{y}}_t$ from (2.4) and (5.2), it follows that

$$\begin{aligned}
&\mathbf{z}_{k\eta} - \hat{\mathbf{y}}_{k\eta} \\
&= \mathbf{z}_{(k-1)\eta} - \hat{\mathbf{y}}_{(k-1)\eta} + \int_{(k-1)\eta}^{k\eta} f(T-t) \left(\mathbf{z}_{(k-1)\eta} - \hat{\mathbf{y}}_{(k-1)\eta}\right) dt \\
&\qquad + \int_{(k-1)\eta}^{k\eta} (g(T-t))^2 \left(\nabla \log p_{T-t}(\mathbf{z}_{(k-1)\eta}) - \nabla \log p_{T-t}\left(\hat{\mathbf{y}}_{(k-1)\eta}\right)\right) dt \\
&\quad + \int_{(k-1)\eta}^{k\eta} \Big[f(T-t)(\mathbf{z}_t - \mathbf{z}_{(k-1)\eta}) \\
&\qquad\qquad + (g(T-t))^2 \left(\nabla \log p_{T-t}(\mathbf{z}_t) - \nabla \log p_{T-t}(\mathbf{z}_{(k-1)\eta})\right) \Big] dt \\
&\qquad + \int_{(k-1)\eta}^{k\eta} (g(T-t))^2 \left(\nabla \log p_{T-t}\left(\hat{\mathbf{y}}_{(k-1)\eta}\right) - s_\theta\left(\hat{\mathbf{y}}_{(k-1)\eta}, T-(k-1)\eta\right)\right) dt.
\end{aligned}$$

This implies that

$$
\begin{aligned}
&\left\|\mathbf{z}_{k\eta}-\hat{\mathbf{y}}_{k\eta}\right\|_{L_2} \\
&\leq \Bigg\|\mathbf{z}_{(k-1)\eta}-\hat{\mathbf{y}}_{(k-1)\eta}+\int_{(k-1)\eta}^{k\eta} f(T-t)\left(\mathbf{z}_{(k-1)\eta}-\hat{\mathbf{y}}_{(k-1)\eta}\right) dt \\
&\qquad +\int_{(k-1)\eta}^{k\eta}(g(T-t))^2\left(\nabla \log p_{T-t}(\mathbf{z}_{(k-1)\eta})-\nabla \log p_{T-t}\left(\hat{\mathbf{y}}_{(k-1)\eta}\right)\right) dt\Bigg\|_{L_2} \\
&+\Bigg\|\int_{(k-1)\eta}^{k\eta}\Big[f(T-t)(\mathbf{z}_t-\mathbf{z}_{(k-1)\eta}) \\
&\qquad\qquad +(g(T-t))^2\left(\nabla \log p_{T-t}(\mathbf{z}_t)-\nabla \log p_{T-t}(\mathbf{z}_{(k-1)\eta})\right)\Big] dt\Bigg\|_{L_2} \\
&\quad +\left\|\int_{(k-1)\eta}^{k\eta}(g(T-t))^2\left(\nabla \log p_{T-t}\left(\hat{\mathbf{y}}_{(k-1)\eta}\right)-s_\theta\left(\hat{\mathbf{y}}_{(k-1)\eta}, T-(k-1)\eta\right)\right) dt\right\|_{L_2}. \qquad (5.6)
\end{aligned}
$$

Next, we provide upper bounds for the three terms in (5.6).

Bounding the first term in (5.6)**.** We can compute that

$$
\begin{aligned}
&\Bigg\|\mathbf{z}_{(k-1)\eta}-\hat{\mathbf{y}}_{(k-1)\eta}+\int_{(k-1)\eta}^{k\eta} f(T-t)\left(\mathbf{z}_{(k-1)\eta}-\hat{\mathbf{y}}_{(k-1)\eta}\right) dt \\
&\qquad +\int_{(k-1)\eta}^{k\eta}(g(T-t))^2\left(\nabla \log p_{T-t}(\mathbf{z}_{(k-1)\eta})-\nabla \log p_{T-t}\left(\hat{\mathbf{y}}_{(k-1)\eta}\right)\right) dt\Bigg\|^2 \\
&=\left\|\mathbf{z}_{(k-1)\eta}-\hat{\mathbf{y}}_{(k-1)\eta}\right\|^2+\Bigg\|\int_{(k-1)\eta}^{k\eta} f(T-t)\left(\mathbf{z}_{(k-1)\eta}-\hat{\mathbf{y}}_{(k-1)\eta}\right) dt \\
&\qquad +\int_{(k-1)\eta}^{k\eta}(g(T-t))^2\left(\nabla \log p_{T-t}(\mathbf{z}_{(k-1)\eta})-\nabla \log p_{T-t}\left(\hat{\mathbf{y}}_{(k-1)\eta}\right)\right) dt\Bigg\|^2 \\
&\qquad\quad +2\int_{(k-1)\eta}^{k\eta} f(T-t)\left\|\mathbf{z}_{(k-1)\eta}-\hat{\mathbf{y}}_{(k-1)\eta}\right\|^2 dt \\
&\quad +2\int_{(k-1)\eta}^{k\eta}\Bigg\langle \mathbf{z}_{(k-1)\eta}-\hat{\mathbf{y}}_{(k-1)\eta}, \\
&\qquad\qquad (g(T-t))^2\left(\nabla \log p_{T-t}(\mathbf{z}_{(k-1)\eta})-\nabla \log p_{T-t}\left(\hat{\mathbf{y}}_{(k-1)\eta}\right)\right)\Bigg\rangle dt.
\end{aligned}
$$

From the proof of Proposition 7, we know that $\log p_{T-t}(\mathbf{x})$ is $a(T-t)$-strongly-concave, where $a(T-t)$ is given in (5.4). Hence we have

$$\begin{aligned}
&\left\| \mathbf{z}_{(k-1)\eta} - \hat{\mathbf{y}}_{(k-1)\eta} + \int_{(k-1)\eta}^{k\eta} f(T-t)\left(\mathbf{z}_{(k-1)\eta} - \hat{\mathbf{y}}_{(k-1)\eta}\right) dt \right. \\
&\qquad \left. + \int_{(k-1)\eta}^{k\eta} (g(T-t))^2 \left(\nabla \log p_{T-t}(\mathbf{z}_{(k-1)\eta}) - \nabla \log p_{T-t}\left(\hat{\mathbf{y}}_{(k-1)\eta}\right)\right) dt \right\|^2 \\
&\leq \left(1 - \int_{(k-1)\eta}^{k\eta} m(T-t)dt\right) \left\|\mathbf{z}_{(k-1)\eta} - \hat{\mathbf{y}}_{(k-1)\eta}\right\|^2 \\
&\qquad + \left(\int_{(k-1)\eta}^{k\eta} f(T-t) \left\|\mathbf{z}_{(k-1)\eta} - \hat{\mathbf{y}}_{(k-1)\eta}\right\| dt \right. \\
&\qquad\qquad \left. + \int_{(k-1)\eta}^{k\eta} (g(T-t))^2 L(T-t) \left\|\mathbf{z}_{(k-1)\eta} - \hat{\mathbf{y}}_{(k-1)\eta}\right\| dt \right)^2 \\
&\leq \left(1 - \int_{(k-1)\eta}^{k\eta} m(T-t)dt + 2\eta \int_{(k-1)\eta}^{k\eta} (f(T-t))^2 dt \right. \\
&\qquad\qquad \left. + 2\eta \int_{(k-1)\eta}^{k\eta} (g(T-t))^4 (L(T-t))^2 dt \right) \cdot \left\|\mathbf{z}_{(k-1)\eta} - \hat{\mathbf{y}}_{(k-1)\eta}\right\|^2,
\end{aligned}$$

where we applied Cauchy-Schwartz inequality and Lemma 9, and $m(T-t)$ is defined in (3.11). Hence, we conclude that

$$\begin{aligned}
&\left\| \mathbf{z}_{(k-1)\eta} - \hat{\mathbf{y}}_{(k-1)\eta} + \int_{(k-1)\eta}^{k\eta} f(T-t)\left(\mathbf{z}_{(k-1)\eta} - \hat{\mathbf{y}}_{(k-1)\eta}\right) dt \right. \\
&\qquad \left. + \int_{(k-1)\eta}^{k\eta} (g(T-t))^2 \left(\nabla \log p_{T-t}\left(\mathbf{z}_{(k-1)\eta}\right) - \nabla \log p_{T-t}\left(\hat{\mathbf{y}}_{(k-1)\eta}\right)\right) dt \right\|_{L_2} \\
&\leq \left(1 - \int_{(k-1)\eta}^{k\eta} \mu(T-t)dt\right) \left\|\mathbf{z}_{(k-1)\eta} - \hat{\mathbf{y}}_{(k-1)\eta}\right\|_{L_2}, \qquad (5.7)
\end{aligned}$$

where we used the inequality $\sqrt{1-x} \leq 1 - \frac{x}{2}$ for any $0 \leq x \leq 1$ and the definition of $\mu(T-t)$ in (3.10) which can be rewritten as

$$\mu(T-t) := (g(T-t))^2 a(T-t) - f(T-t) - \eta(f(T-t))^2 - \eta(g(T-t))^4 (L(T-t))^2, \quad 0 \leq t \leq T,$$

where $a(T-t)$ is given in (5.4).

Bounding the second term in (5.6)**.** Using Lemma 9, we can compute that

$$\begin{aligned}
&\left\|\int_{(k-1)\eta}^{k\eta}\left[f(T-t)(\mathbf{z}_t-\mathbf{z}_{(k-1)\eta})+(g(T-t))^2\left(\nabla\log p_{T-t}(\mathbf{z}_t)-\nabla\log p_{T-t}(\mathbf{z}_{(k-1)\eta})\right)\right]dt\right\|^2\\
&\le\left(\int_{(k-1)\eta}^{k\eta}[f(T-t)+(g(T-t))^2L(T-t)]\|\mathbf{z}_t-\mathbf{z}_{(k-1)\eta}\|dt\right)^2\\
&\le\eta\int_{(k-1)\eta}^{k\eta}[f(T-t)+(g(T-t))^2L(T-t)]^2\|\mathbf{z}_t-\mathbf{z}_{(k-1)\eta}\|^2dt,
\end{aligned}$$

which implies that

$$\begin{aligned}
&\left\|\int_{(k-1)\eta}^{k\eta}\left[f(T-t)(\mathbf{z}_t-\mathbf{z}_{(k-1)\eta})+(g(T-t))^2\left(\nabla\log p_{T-t}(\mathbf{z}_t)-\nabla\log p_{T-t}\left(\mathbf{z}_{(k-1)\eta}\right)\right)\right]dt\right\|_{L_2}\\
&\le\left(\eta\int_{(k-1)\eta}^{k\eta}[f(T-t)+(g(T-t))^2L(T-t)]^2dt\cdot\sup_{(k-1)\eta\le t\le k\eta}\mathbb{E}\|\mathbf{z}_t-\mathbf{z}_{(k-1)\eta}\|^2\right)^{1/2}\\
&=\sqrt{\eta}\left(\int_{(k-1)\eta}^{k\eta}[f(T-t)+(g(T-t))^2L(T-t)]^2dt\right)^{1/2}\sup_{(k-1)\eta\le t\le k\eta}\left\|\mathbf{z}_t-\mathbf{z}_{(k-1)\eta}\right\|_{L_2}.
\end{aligned}\tag{5.8}$$

Bounding the third term in (5.6)**.** We notice that

$$\begin{aligned}
&\left\|\int_{(k-1)\eta}^{k\eta}(g(T-t))^2\left(\nabla\log p_{T-t}(\hat{\mathbf{y}}_{(k-1)\eta})-s_\theta\left(\hat{\mathbf{y}}_{(k-1)\eta},T-(k-1)\eta\right)\right)dt\right\|_{L_2}\\
&\le\left\|\int_{(k-1)\eta}^{k\eta}(g(T-t))^2\left(\nabla\log p_{T-(k-1)\eta}(\hat{\mathbf{y}}_{(k-1)\eta})-s_\theta\left(\hat{\mathbf{y}}_{(k-1)\eta},T-(k-1)\eta\right)\right)dt\right\|_{L_2}\\
&\quad+\left\|\int_{(k-1)\eta}^{k\eta}(g(T-t))^2\left(\nabla\log p_{T-t}\left(\hat{\mathbf{y}}_{(k-1)\eta}\right)-\nabla\log p_{T-(k-1)\eta}\left(\hat{\mathbf{y}}_{(k-1)\eta}\right)\right)dt\right\|_{L_2}.
\end{aligned}$$

By Assumption 3, we have

$$\begin{aligned}
&\left\|\int_{(k-1)\eta}^{k\eta}(g(T-t))^2\left(\nabla\log p_{T-(k-1)\eta}(\hat{\mathbf{y}}_{(k-1)\eta})-s_\theta\left(\hat{\mathbf{y}}_{(k-1)\eta},T-(k-1)\eta\right)\right)dt\right\|_{L_2}\\
&\le M\int_{(k-1)\eta}^{k\eta}(g(T-t))^2dt.
\end{aligned}\tag{5.9}$$

Moreover, by Assumption 2, we have

$$\begin{aligned}
&\left\| \int_{(k-1)\eta}^{k\eta} (g(T-t))^2 \left(\nabla \log p_{T-t}\left(\hat{\mathbf{y}}_{(k-1)\eta}\right) - \nabla \log p_{T-(k-1)\eta}\left(\hat{\mathbf{y}}_{(k-1)\eta}\right)\right) dt \right\|_{L_2} \\
&\leq \int_{(k-1)\eta}^{k\eta} (g(T-t))^2 \left\| \nabla \log p_{T-t}\left(\hat{\mathbf{y}}_{(k-1)\eta}\right) - \nabla \log p_{T-(k-1)\eta}\left(\hat{\mathbf{y}}_{(k-1)\eta}\right)\right\|_{L_2} dt \\
&\leq \int_{(k-1)\eta}^{k\eta} (g(T-t))^2 M_1 \eta \left(1 + \left\|\hat{\mathbf{y}}_{(k-1)\eta}\right\|_{L_2}\right) dt \\
&\leq M_1 \eta \left(1 + \left\|\mathbf{z}_{(k-1)\eta} - \hat{\mathbf{y}}_{(k-1)\eta}\right\|_{L_2} + \|\mathbf{z}_{(k-1)\eta}\|_{L_2}\right) \int_{(k-1)\eta}^{k\eta} (g(T-t))^2 dt. \qquad (5.10)
\end{aligned}$$

Furthermore, we can compute that

$$\left\|\mathbf{z}_{(k-1)\eta}\right\|_{L_2} \leq \left\|\mathbf{z}_{(k-1)\eta} - \tilde{\mathbf{x}}_{(k-1)\eta}\right\|_{L_2} + \left\|\tilde{\mathbf{x}}_{(k-1)\eta}\right\|_{L_2}, \qquad (5.11)$$

where $\tilde{\mathbf{x}}_t$ is defined in (1.2). Moreover, by the proof of Proposition 7, we have

$$\left\|\mathbf{z}_{(k-1)\eta} - \tilde{\mathbf{x}}_{(k-1)\eta}\right\|_{L_2} \leq \left(\mathbb{E}\|\tilde{\mathbf{x}}_0 - \mathbf{z}_0\|^2\right)^{1/2} = e^{-\int_0^T f(s)ds} \|\mathbf{x}_0\|_{L_2} \leq \|\mathbf{x}_0\|_{L_2}, \qquad (5.12)$$

where we applied (2.1) to obtain the equality in the above equation. Moreover, since $(\tilde{\mathbf{x}}_t)_{0\leq t\leq T}$ is the time-reversal process of $(\mathbf{x}_t)_{0\leq t\leq T}$, we have

$$\left\|\tilde{\mathbf{x}}_{(k-1)\eta}\right\|_{L_2} = \left\|\mathbf{x}_{T-(k-1)\eta}\right\|_{L_2} \leq \sup_{0\leq t\leq T} \|\mathbf{x}_t\|_{L_2} =: c_2(T). \qquad (5.13)$$

Next, let us show that $c_2(T)$ can be computed as given by the formula in (3.13). By applying Itô's formula to equation (2.1), we have

$$\begin{aligned}
&d\left(\|\mathbf{x}_t\|^2 e^{2\int_0^t f(s)ds}\right) \\
&= 2f(t)\|\mathbf{x}_t\|^2 e^{2\int_0^t f(s)ds} dt + 2e^{2\int_0^t f(s)ds} \langle \mathbf{x}_t, d\mathbf{x}_t \rangle + e^{2\int_0^t f(s)ds} \cdot d \cdot (g(t))^2 dt.
\end{aligned}$$

By taking expectations, we obtain

$$d\left(\mathbb{E}\|\mathbf{x}_t\|^2 e^{2\int_0^t f(s)ds}\right) = e^{2\int_0^t f(s)ds} \cdot d \cdot (g(t))^2 dt,$$

so that

$$\mathbb{E}\|\mathbf{x}_t\|^2 = e^{-2\int_0^t f(s)ds} \mathbb{E}\|\mathbf{x}_0\|^2 + d\int_0^t e^{-2\int_s^t f(v)dv} (g(s))^2 ds.$$

Therefore, we conclude that

$$c_2(T) = \sup_{0\leq t\leq T} \|\mathbf{x}_t\|_{L_2} = \sup_{0\leq t\leq T} \left(e^{-2\int_0^t f(s)ds} \|\mathbf{x}_0\|_{L_2}^2 + d\int_0^t e^{-2\int_s^t f(v)dv} (g(s))^2 ds \right)^{1/2}. \qquad (5.14)$$

Therefore, by applying (5.10), (5.11), (5.12) and (5.13), we have

$$\begin{aligned}&\left\|\int_{(k-1)\eta}^{k\eta}(g(T-t))^2\left(\nabla\log p_{T-t}\left(\hat{\mathbf{y}}_{(k-1)\eta}\right)-\nabla\log p_{T-(k-1)\eta}\left(\hat{\mathbf{y}}_{(k-1)\eta}\right)\right)dt\right\|_{L_2}\\&\leq\left(M_1\eta\left\|\mathbf{z}_{(k-1)\eta}-\hat{\mathbf{y}}_{(k-1)\eta}\right\|_{L_2}+M_1\eta\left(1+\|\mathbf{x}_0\|_{L_2}+c_2(T)\right)\right)\int_{(k-1)\eta}^{k\eta}(g(T-t))^2dt.\end{aligned}$$

It follows that the third term in (5.6) is upper bounded by

$$\left(M_1\eta\left\|\mathbf{z}_{(k-1)\eta}-\hat{\mathbf{y}}_{(k-1)\eta}\right\|_{L_2}+M_1\eta\left(1+\|\mathbf{x}_0\|_{L_2}+c_2(T)\right)+M\right)\int_{(k-1)\eta}^{k\eta}(g(T-t))^2dt. \tag{5.15}$$

Bounding (5.6)**.** On combining (5.7), (5.8) and (5.15), we conclude that

$$\begin{aligned}&\|\mathbf{z}_{k\eta}-\hat{\mathbf{y}}_{k\eta}\|_{L_2}^2 \qquad (5.16)\\&\leq\Bigg\{\gamma_{k,\eta}\left\|\mathbf{z}_{(k-1)\eta}-\hat{\mathbf{y}}_{(k-1)\eta}\right\|_{L_2}\\&\quad+M_1\eta\left(1+\|\mathbf{x}_0\|_{L_2}+c_2(T)\right)\int_{(k-1)\eta}^{k\eta}(g(T-t))^2dt+M\int_{(k-1)\eta}^{k\eta}(g(T-t))^2dt\\&\quad+\sqrt{\eta}\left(\int_{(k-1)\eta}^{k\eta}[f(T-t)+(g(T-t))^2L(T-t)]^2dt\right)^{1/2}\sup_{(k-1)\eta\leq t\leq k\eta}\|\mathbf{z}_t-\mathbf{z}_{(k-1)\eta}\|_{L_2}\Bigg\}^2,\end{aligned}$$

where we used the definition of $\gamma_{k,\eta}$ in (3.14). We need one more result, which provides an upper bound for $\sup_{(k-1)\eta\leq t\leq k\eta}\|\mathbf{z}_t-\mathbf{z}_{(k-1)\eta}\|_{L_2}$. The proof of Lemma 10 is given in Appendix A.2.

Lemma 10 *For any $k=1,2,\ldots,K$,*

$$\begin{aligned}\sup_{(k-1)\eta\leq t\leq k\eta}\left\|\mathbf{z}_t-\mathbf{z}_{(k-1)\eta}\right\|_{L_2}&\leq c_1(T)\int_{(k-1)\eta}^{k\eta}\left[f(T-s)+(g(T-s))^2L(T-s)\right]ds\\&\quad+c_2(T)\int_{T-k\eta}^{T-(k-1)\eta}f(s)ds+\left(\int_{T-k\eta}^{T-(k-1)\eta}(g(s))^2ds\right)^{1/2}\sqrt{d},\end{aligned}$$

where $c_1(T)$ and $c_2(T)$ are given in (3.12)-(3.13) *respectively.*

By applying Lemma 10, we conclude from (5.16) that

$$\begin{aligned}
&\|\mathbf{z}_{k\eta} - \hat{\mathbf{y}}_{k\eta}\|_{L_2} \\
&\leq \gamma_{k,\eta} \left\|\mathbf{z}_{(k-1)\eta} - \hat{\mathbf{y}}_{(k-1)\eta}\right\|_{L_2} \\
&\quad + M_1\eta\left(1 + \|\mathbf{x}_0\|_{L_2} + c_2(T)\right)\int_{(k-1)\eta}^{k\eta}(g(T-t))^2dt + M\int_{(k-1)\eta}^{k\eta}(g(T-t))^2dt \\
&\quad + \sqrt{\eta}h_{k,\eta}\left(\int_{(k-1)\eta}^{k\eta}[f(T-t) + (g(T-t))^2L(T-t)]^2dt\right)^{1/2},
\end{aligned}$$

where $h_{k,\eta}$ is defined in (3.15). The proof of Proposition 8 is hence complete. ■

5.2 Proof of Proposition 5

Proof By (3.15), we have $h_{k,\eta} \geq \min_{0\leq t\leq T} g(t)\sqrt{\eta}\sqrt{d}$. Therefore, we have

$$\begin{aligned}
&\text{RHS of (3.16)} \\
&\geq e^{-\int_0^{K\eta}c(t)dt}\|\mathbf{x}_0\|_{L_2} + \sum_{k=1}^{K}\prod_{j=k+1}^{K}\left(1 - \int_{(j-1)\eta}^{j\eta}\mu(T-t)dt\right) \\
&\quad \cdot \left(M\int_{(k-1)\eta}^{k\eta}(g(T-t))^2dt + \sqrt{\eta}h_{k,\eta}\left(\int_{(k-1)\eta}^{k\eta}[f(T-t) + (g(T-t))^2L(T-t)]^2dt\right)^{1/2}\right) \\
&\geq e^{-\int_0^{K\eta}c(t)dt}\|\mathbf{x}_0\|_{L_2} \\
&\quad + \sum_{k=1}^{K}\left(1 - \eta\max_{0\leq t\leq T}\mu(t)\right)^{K-k}\left(\sqrt{\eta}\min_{0\leq t\leq T}g(t)\sqrt{\eta}\sqrt{d}\sqrt{\eta}\min_{0\leq t\leq T}\left(f(t) + (g(t))^2L(t)\right)\right) \\
&\geq e^{-\int_0^{K\eta}c(t)dt}\|\mathbf{x}_0\|_{L_2} \\
&\quad + \sqrt{\eta}\sqrt{d}\frac{1 - e^{-K\eta\max_{0\leq t\leq T}\mu(t)}}{\max_{0\leq t\leq T}\mu(t)}\left(\min_{0\leq t\leq T}g(t)\min_{0\leq t\leq T}\left(f(t) + (g(t))^2L(t)\right)\right),
\end{aligned}$$

where the equality above is due to the formula for the finite sum of a geometric series and we used the inequality that $1 - x \leq e^{-x}$ for any $0 \leq x \leq 1$ to obtain the last inequality above. Therefore, in order for RHS of (3.16) $\leq \epsilon$, we must have $e^{-\int_0^{K\eta}c(t)dt}\|\mathbf{x}_0\|_{L_2} \leq \epsilon$, which implies that $K\eta \to \infty$ as $\epsilon \to 0$ and in particular

$$T = K\eta = \Omega(1), \tag{5.17}$$

and we also need

$$\sqrt{\eta}\sqrt{d}\frac{1 - e^{-K\eta\max_{0\leq t\leq T}\mu(t)}}{\max_{0\leq t\leq T}\mu(t)}\left(\min_{0\leq t\leq T}g(t)\min_{0\leq t\leq T}\left(f(t) + (g(t))^2L(t)\right)\right) \leq \epsilon. \tag{5.18}$$

Note that by the definition of $\mu(t)$ in (3.11) and $c(t)$ in (3.9), we have $\max_{0\le t\le T}\mu(t)\le \max_{0\le t\le T}c(t)=\max_{0\le t\le K\eta}c(t)$. Note Assumption 1 implies that

$$\|\mathbf{x}_0\|_{L_2}\le\sqrt{2d/m_0}+\|\mathbf{x}_*\|, \tag{5.19}$$

where $\mathbf{x}_*$ is the unique minimizer of $-\log p_0$. See Lemma 11 in Gürbüzbalaban et al. (2021). Hence we have $e^{-\int_0^{K\eta}c(t)dt}=\mathcal{O}(\epsilon/\sqrt{d})$. Together with the assumption that $\max_{0\le s\le t}c(s)\le c_1\left(\int_0^t c(s)ds\right)^{\rho}+c_2$ uniformly in t for some $c_1,c_2,\rho>0$, it is easy to see that $\max_{0\le t\le K\eta}c(t)=\mathcal{O}\left(\left(\log\left(\sqrt{d}/\epsilon\right)\right)^{\rho}\right)$ and hence $\max_{0\le t\le T}\mu(t)=\mathcal{O}\left(\left(\log\left(\sqrt{d}/\epsilon\right)\right)^{\rho}\right)$. Moreover, under our assumption $\mu(t)\ge\frac{1}{4}m(t)$ for any $0\le t\le T$, where $\mu(t),m(t)$ are defined in (3.10) and (3.11) and since we assumed $\min_{t\ge0}g(t)>0$, we have $m(t)>0$ for any t. Together with $T=K\eta=\Omega(1)$ from (5.17), we have $\max_{0\le t\le T}\mu(t)\ge\Omega(1)$. Since $K\eta\to\infty$ as $\epsilon\to0$, we have $1-e^{-K\eta\max_{0\le t\le T}\mu(t)}=\Omega(1)$. Therefore, it follows from (5.18) that $\eta=\widetilde{\mathcal{O}}\left(\frac{\epsilon^2}{d}\right)$, where $\widetilde{\mathcal{O}}$ ignores the logarithmic dependence on ϵ and d and we used the assumptions that $\min_{t\ge0}g(t)>0$ and $\min_{t\ge0}(f(t)+(g(t))^2L(t))>0$. Hence, we conclude that we have the following lower bound for the complexity: $K=\widetilde{\Omega}\left(\frac{d}{\epsilon^2}\right)$, where $\widetilde{\Omega}$ ignores the logarithmic dependence on ϵ and d. This completes the proof. ■

5.3 Proof of Proposition 6

Proof When $\mathbf{x}_0\sim\mathcal{N}(0,\sigma_0^2I_d)$, we can compute that

$$\begin{aligned}
\nabla_{\mathbf{x}}\log p_{T-t}(\mathbf{x}) &= \nabla_{\mathbf{x}}\log\int_{\mathbb{R}^d}\exp\left(-\frac{\|\mathbf{x}-e^{-\int_0^{T-t}f(s)ds}\mathbf{x}_0\|^2}{2\int_0^{T-t}e^{-2\int_s^{T-t}f(v)dv}(g(s))^2ds}\right)p_0(\mathbf{x}_0)d\mathbf{x}_0\\
&= \nabla_{\mathbf{x}}\log\int_{\mathbb{R}^d}\exp\left(-\frac{\|\mathbf{x}-e^{-\int_0^{T-t}f(s)ds}\mathbf{x}_0\|^2}{2\int_0^{T-t}e^{-2\int_s^{T-t}f(v)dv}(g(s))^2ds}\right)\exp\left(-\frac{\|\mathbf{x}_0\|^2}{2\sigma_0^2}\right)d\mathbf{x}_0\\
&= \nabla_{\mathbf{x}}\log\int_{\mathbb{R}^d}\exp\left(-\frac{\|((a_1(T-t))^2\sigma_0^2+a_2(T-t))\mathbf{x}_0-a_1(T-t)\sigma_0^2\mathbf{x}\|^2}{2a_2(T-t)\sigma_0^2((a_1(T-t))^2\sigma_0^2+a_2(T-t))}\right)\\
&\qquad\cdot\exp\left(-\frac{\|\mathbf{x}\|^2}{2((a_1(T-t))^2\sigma_0^2+a_2(T-t))}\right)d\mathbf{x}_0\\
&= -\frac{1}{(a_1(T-t))^2\sigma_0^2+a_2(T-t)}\mathbf{x},
\end{aligned}$$

where

$$a_1(T-t):=e^{-\int_0^{T-t}f(s)ds},\qquad a_2(T-t):=\int_0^{T-t}e^{-2\int_s^{T-t}f(v)dv}(g(s))^2ds. \tag{5.20}$$

Therefore, under the assumption $\mathbf{x}_0\sim\mathcal{N}(0,\sigma_0^2I_d)$, the discretization of (2.4) is given by:

$$\mathbf{y}_k=\left(1-\int_{(k-1)\eta}^{k\eta}\alpha(T-t)dt\right)\mathbf{y}_{k-1}+\left(\int_{(k-1)\eta}^{k\eta}(g(T-t))^2dt\right)^{1/2}\xi_k, \tag{5.21}$$

where

$$\alpha(T-t) := \frac{(g(T-t))^2}{(a_1(T-t))^2\sigma_0^2 + a_2(T-t)} - f(T-t), \tag{5.22}$$

where $a_1(T-t)$ and $a_2(T-t)$ are defined in (5.20), and ξ_k are i.i.d. Gaussian random vectors $\mathcal{N}(0, I_d)$ and $\mathbf{y}_0$ follows the same distribution as $\hat{p}_T$ in (2.2).

Since $\mathbf{y}_K$ and $\mathbf{x}_0$ are both centered Gaussian random vectors, by using the explicit formula for the $\mathcal{W}_2$ distance between two Gaussian distributions, we have

$$\mathcal{W}_2\left(\mathcal{L}(\mathbf{y}_K), \mathcal{L}(\mathbf{x}_0)\right) = \left(\mathrm{Tr}\left(\Sigma_K + \sigma_0^2 I_d - 2\left(\Sigma_K^{1/2}\sigma_0^2 I_d \Sigma_K^{1/2}\right)^{1/2}\right)\right)^{1/2},$$

where $\Sigma_k = \mathbb{E}\left[\mathbf{y}_k \mathbf{y}_k^\top\right]$, $k = 0, 1, \ldots, K$, is the covariance matrix of $\mathbf{y}_k$. It is easy to compute that for any $k = 1, 2, \ldots, K$, $\Sigma_k = \left(1 - \int_{(k-1)\eta}^{k\eta} \alpha(T-t)dt\right)^2 \Sigma_{k-1} + \int_{(k-1)\eta}^{k\eta} (g(T-t))^2 dt \cdot I_d$, where $\alpha(T-t)$ is defined in (5.22) with $\Sigma_0 = \int_0^T e^{-2\int_s^T f(v)dv}(g(s))^2 ds \cdot I_d$. Therefore, one can deduce that $\Sigma_k = \hat{\sigma}_k^2 I_d$, $k = 0, 1, \ldots, K$, where

$$\hat{\sigma}_k^2 = \left(1 - \int_{(k-1)\eta}^{k\eta} \alpha(T-t)dt\right)^2 \hat{\sigma}_{k-1}^2 + \int_{(k-1)\eta}^{k\eta} (g(T-t))^2 dt, \tag{5.23}$$

with $\hat{\sigma}_0^2 = \int_0^T e^{-2\int_s^T f(v)dv}(g(s))^2 ds$. Moreover, we get

$$\mathcal{W}_2\left(\mathcal{L}(\mathbf{y}_K), \mathcal{L}(\mathbf{x}_0)\right) = \left(\mathrm{Tr}\left(\hat{\sigma}_K^2 I_d + \sigma_0^2 I_d - 2\left(\hat{\sigma}_K \sigma_0^2 \hat{\sigma}_K\right)^{1/2} I_d\right)\right)^{1/2} = \sqrt{d}\left|\hat{\sigma}_K - \sigma_0\right|. \tag{5.24}$$

One can easily compute from (5.23) that

$$\begin{aligned}\hat{\sigma}_K^2 = \sum_{j=1}^K \prod_{i=j+1}^K \left(1 - \int_{(i-1)\eta}^{i\eta} \alpha(T-t)dt\right)^2 \int_{(j-1)\eta}^{j\eta} (g(T-t))^2 dt \\ + \prod_{i=1}^K \left(1 - \int_{(i-1)\eta}^{i\eta} \alpha(T-t)dt\right)^2 \int_0^T e^{-2\int_s^T f(v)dv}(g(s))^2 ds,\end{aligned} \tag{5.25}$$

where $\alpha(T-t)$ is defined in (5.22).

By discrete approximation of a Riemann integral, with fixed $K\eta = T$, we have

$$\left|\hat{\sigma}_K^2 - \int_0^T e^{-2\int_t^T \alpha(T-s)ds}(g(T-t))^2 dt - e^{-2\int_0^T \alpha(s)ds} \int_0^T e^{-2\int_s^T f(v)dv}(g(s))^2 ds\right| = \Theta(\eta),$$

as $\eta \to 0$. Indeed, one can show that there exists some $c_0 \in \mathbb{R}$, such that

$$\hat{\sigma}_K^2 = \int_0^T e^{-2\int_t^T \alpha(T-s)ds}(g(T-t))^2 dt + e^{-2\int_0^T \alpha(s)ds} \int_0^T e^{-2\int_s^T f(v)dv}(g(s))^2 ds + c_0\eta + \mathcal{O}(\eta^2), \tag{5.26}$$

as $\eta \to 0$. Next, let us show that (5.26) holds as well as spell out the constant c_0 explicitly.

First, we can compute that

$$\log\prod_{i=1}^{K}\left(1-\int_{(i-1)\eta}^{i\eta}\alpha(T-t)dt\right)^2 = 2\sum_{i=1}^{K}\log\left(1-\int_{(i-1)\eta}^{i\eta}\alpha(T-t)dt\right)$$
$$= 2\sum_{i=1}^{K}\left(-\int_{(i-1)\eta}^{i\eta}\alpha(T-t)dt-\frac{1}{2}\left(\int_{(i-1)\eta}^{i\eta}\alpha(T-t)dt\right)^2+\mathcal{O}(\eta^3)\right),$$

which implies that

$$\log\prod_{i=1}^{K}\left(1-\int_{(i-1)\eta}^{i\eta}\alpha(T-t)dt\right)^2-\log\left(e^{-2\int_0^T\alpha(T-t)dt}\right)=-\eta\int_0^T(\alpha(T-t))^2dt+\mathcal{O}(\eta^2).$$

Therefore, we have

$$\prod_{i=1}^{K}\left(1-\int_{(i-1)\eta}^{i\eta}\alpha(T-t)dt\right)^2 = e^{-2\int_0^T\alpha(s)ds}e^{\log\prod_{i=1}^{K}\left(1-\int_{(i-1)\eta}^{i\eta}\alpha(T-t)dt\right)^2-\log\left(e^{-2\int_0^T\alpha(T-t)dt}\right)}$$
$$= e^{-2\int_0^T\alpha(s)ds}\left(1-\eta\int_0^T(\alpha(T-t))^2dt+\mathcal{O}(\eta^2)\right).$$

Similarly, we can show that

$$\sum_{j=1}^{K}\prod_{i=j+1}^{K}\left(1-\int_{(i-1)\eta}^{i\eta}\alpha(T-t)dt\right)^2\int_{(j-1)\eta}^{j\eta}(g(T-t))^2dt$$
$$=\sum_{j=1}^{K}\int_{(j-1)\eta}^{j\eta}e^{-2\int_{j\eta}^{K\eta}\alpha(T-s)ds}(g(T-t))^2dt$$
$$-\eta\int_0^T e^{-2\int_t^T\alpha(T-s)ds}\left(\int_t^T(\alpha(T-s))^2ds\right)(g(T-t))^2dt+\mathcal{O}(\eta^2).$$

Moreover,

$$\sum_{j=1}^{K}\int_{(j-1)\eta}^{j\eta}e^{-2\int_{j\eta}^{K\eta}\alpha(T-s)ds}(g(T-t))^2dt$$
$$=\int_0^T e^{-2\int_t^T\alpha(T-s)ds}(g(T-t))^2dt+\eta\int_0^T e^{-2\int_t^T\alpha(T-s)ds}\alpha(T-t)(g(T-t))^2dt+\mathcal{O}(\eta^2).$$

Hence, we conclude that (5.26) holds with

$$c_0=-e^{-2\int_0^T\alpha(s)ds}\int_0^T(\alpha(t))^2dt\int_0^T e^{-2\int_s^T f(v)dv}(g(s))^2ds$$
$$-\int_0^T e^{-2\int_0^t\alpha(s)ds}\left(\int_0^t(\alpha(s))^2ds\right)(g(t))^2dt+\int_0^T e^{-2\int_0^t\alpha(s)ds}\alpha(t)(g(t))^2dt. \quad (5.27)$$

On the other hand, $\tilde{\mathbf{x}}_{T-t}$ has the same distribution as $\mathbf{x}_t$ for any $0 \le t \le T$, where

$$d\tilde{\mathbf{x}}_t = -\alpha(T-t)\tilde{\mathbf{x}}_t dt + g(T-t)d\bar{\mathbf{B}}_t,$$

with $\tilde{\mathbf{x}}_0 \sim p_T$ such that

$$\tilde{\mathbf{x}}_0 = e^{-\int_0^T f(s)ds}\mathbf{x}_0 + \int_0^T e^{-\int_s^T f(v)dv} g(s)d\mathbf{B}_s.$$

Since $\tilde{\mathbf{x}}_T$ has the same distribution as $\mathbf{x}_0$, their covariance matrices are the same such that

$$\begin{aligned}\mathbb{E}\left[\tilde{\mathbf{x}}_T\tilde{\mathbf{x}}_T^\top\right] &= \int_0^T e^{-2\int_t^T \alpha(T-s)ds}(g(T-t))^2 dt \cdot I_d \\ &+ e^{-2\int_0^T \alpha(s)ds}\left(e^{-2\int_0^T f(s)ds}\sigma_0^2 + \int_0^T e^{-2\int_s^T f(v)dv}(g(s))^2 ds\right)\cdot I_d = \mathbb{E}\left[\mathbf{x}_0\mathbf{x}_0^\top\right] = \sigma_0^2 I_d.\end{aligned}$$

This implies that

$$\hat{\sigma}_K^2 = \sigma_0^2 - e^{-2\int_0^T \alpha(s)ds}e^{-2\int_0^T f(s)ds}\sigma_0^2 + c_0\eta + \mathcal{O}(\eta^2), \tag{5.28}$$

where $\alpha(\cdot)$ is defined in (5.22) and c_0 is given in (5.27), and by the definition of $\alpha(\cdot)$, we can equivalently write (5.28) as

$$\hat{\sigma}_K^2 = \sigma_0^2 - e^{-2\int_0^T \gamma(s)ds}\sigma_0^2 + c_0\eta + \mathcal{O}(\eta^2), \tag{5.29}$$

where

$$\gamma(t) := \frac{(g(t))^2}{(a_1(t))^2\sigma_0^2 + a_2(t)}, \tag{5.30}$$

with $a_1(\cdot), a_2(\cdot)$ defined in (5.20).

If there exists some $T = T(\epsilon)$ and $\bar{\eta} = \bar{\eta}(\epsilon)$ such that $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \le \epsilon$ for any $K \ge \bar{K} := T/\bar{\eta}$ (with $\eta = T/K \le \bar{\eta}$) then we must have $\mathcal{W}_2\left(\mathcal{L}(\mathbf{y}_K), \mathcal{L}(\mathbf{x}_0)\right) = \sqrt{d}\,|\hat{\sigma}_K - \sigma_0| \le \epsilon$, for any $K \ge \bar{K} := T/\bar{\eta}$ so that

$$\left|\hat{\sigma}_K^2 - \sigma_0^2\right| = |\hat{\sigma}_K - \sigma_0|(\hat{\sigma}_K + \sigma_0) \le \frac{\epsilon}{\sqrt{d}}\left(2\sigma_0 + \frac{\epsilon}{\sqrt{d}}\right), \tag{5.31}$$

for any $K \ge \bar{K} := T/\bar{\eta}$ (with $\eta = T/K \le \bar{\eta}$). Then, it follows from (5.29) and (5.31) that

$$\left|-e^{-2\int_0^T \gamma(s)ds}\sigma_0^2 + c_0\eta + \mathcal{O}(\eta^2)\right| \le \mathcal{O}\left(\epsilon/\sqrt{d}\right). \tag{5.32}$$

Since (5.32) holds for any $\eta \le \bar{\eta}$, by letting $\eta \to 0$ in (5.32), we get

$$e^{-2\int_0^T \gamma(s)ds}\sigma_0^2 \le \mathcal{O}\left(\epsilon/\sqrt{d}\right). \tag{5.33}$$

By the definition of $\gamma(t)$ in (5.30), it is positive and continuous for any $t > 0$ and therefore, it follows from (5.33) that $T \to \infty$ as $\epsilon \to 0$, and in particular $T \ge \Omega(1)$. Finally, by letting $\eta = \bar{\eta}$ in (5.32), applying (5.33), we deduce that $\bar{\eta} \le \mathcal{O}\left(\frac{\epsilon}{\sqrt{d}}\right)$. Hence, by $T \ge \Omega(1)$, we conclude that $\bar{K} = \frac{T}{\bar{\eta}} \ge \mathcal{O}\left(\frac{\sqrt{d}}{\epsilon}\right)$. This completes the proof. ■

6. Conclusion and Future Work

In this paper, we establish convergence guarantees for a general class of score-based generative models in the 2-Wasserstein distance for smooth log-concave data distributions. Our theoretical result directly leads to iteration complexity bounds for various score-based generative models with different forward processes. Moreover, our experimental results align well with our theoretical predictions on the iteration complexity.

Our work serves as a first step towards a better understanding of the impacts of different choices of forward processes in the SDE implementation of diffusion models. It is a significant open question how to relax the assumption of strong log-concavity on the data distribution. Our convergence analysis borrows the idea of synchronous coupling used in sampling with Langevin algorithms (Dalalyan and Karagulyan, 2019). To go beyond the log-concave setting, one may consider more sophisticated coupling methods such as reflection coupling (see e.g. Eberle (2016)) to obtain contraction rates of SDEs in Wasserstein distance. However, it is not clear whether the reflection coupling can be applied to the convergence analysis of score-based diffusion models, because the reverse SDE is time-inhomogeneous and it is also unclear whether the coefficients in the reverse SDE satisfy a dissipativity-type condition in Eberle (2016). In addition to considering weaker conditions on the target data distribution, another interesting direction is to study the convergence theory for alternative sampling schemes (beyond the Euler-Maruyama discretization of reverse SDEs), such as (stochastic) EDM in Karras et al. (2022). Furthermore, it is also important to study the complexity of score estimations and establish an end-to-end convergence theory for diffusion models (see e.g. Chen et al. (2023b); Han et al. (2024) for some recent progress). Finally, our empirical analysis focuses on image generation using CIFAR-10 data exclusively. It would be intriguing to explore additional datasets and tasks. We leave them for future research.

Acknowledgements

We would like to thank the action editor and two anonymous referees for many constructive comments and suggestions. Xuefeng Gao acknowledges support from the Hong Kong Research Grants Council [GRF 14201421, 14201424, 14200123, 14212522]. Hoang M. Nguyen and Lingjiong Zhu are partially supported by the grants NSF DMS-2053454, NSF DMS-2208303.

References

B. D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

J. Benton, V. De Bortoli, A. Doucet, and G. Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *International Conference on*

Learning Representations, 2024.

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

F. Bolley and C. Villani. Weighted Csiszár-Kullback-pinsker inequalities and applications to transportation inequalities. *Annales-Faculté des sciences Toulouse Mathematiques*, 14 (3):331, 2005.

S. Bruno, Y. Zhang, D.-Y. Lim, Ö. D. Akyildiz, and S. Sabanis. On diffusion-based generative models and their error bounds: The log-concave case with full convergence estimates. *arXiv preprint arXiv:2311.13584*, 2023.

P. Cattiaux, G. Conforti, I. Gentil, and C. Léonard. Time reversal of diffusion processes under a finite entropy condition. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, 59(4):1844–1881, 2023.

H. Chen, H. Lee, and J. Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, volume 202, pages 4764–4803. PMLR, 2023a.

M. Chen, K. Huang, T. Zhao, and M. Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202. PMLR, 2023b.

S. Chen, S. Chewi, H. Lee, Y. Li, J. Lu, and A. Salim. The probability flow ODE is provably fast. In *Advances in Neural Information Processing Systems*, 2023c.

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. In *International Conference on Learning Representations*, 2023d.

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*, volume 202, pages 4462–4484. PMLR, 2023e.

F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:10850–10869, 2023.

A. S. Dalalyan and A. G. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12): 5278–5311, 2019.

V. De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 11:1–42, 2022.

V. De Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. In *Advances in Neural Information Processing Systems*, volume 34, pages 17695–17709, 2021.

A. Eberle. Reflection couplings and contraction rates for diffusions. *Probability Theory and Related Fields*, 166:851–886, 2016.

X. Gao and L. Zhu. Convergence analysis for general probability flow ODEs of diffusion models in Wasserstein distances. *arXiv preprint arXiv:2401.17958*, 2024.

A. Gibbs and F. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70:419–435, 2002.

M. Gürbüzbalaban, X. Gao, Y. Hu, and L. Zhu. Decentralized stochastic gradient Langevin dynamics and Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 22:1–69, 2021.

Y. Han, M. Razaviyayn, and R. Xu. Neural network-based score estimation in diffusion models: Optimization and generalization. In *International Conference on Learning Representations*, 2024.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

A. Hyvärinen and P. Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4):695–708, 2005.

T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

H. Lee, J. Lu, and Y. Tan. Convergence for score-based generative modeling with polynomial complexity. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

H. Lee, J. Lu, and Y. Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, volume 201, pages 946–985. PMLR, 2023.

G. Li, Y. Huang, T. Efimov, Y. Wei, Y. Chi, and Y. Chen. Accelerating convergence of score-based diffusion models, provably. In *International Conference on Machine Learning*, volume 235, pages 27942–27954. PMLR, 2024a.

G. Li, Y. Wei, Y. Chen, and Y. Chi. Towards non-asymptotic convergence for diffusion-based generative models. In *International Conference on Learning Representations*, 2024b.

R. Li, H. Zha, and M. Tao. Sqrt(d) dimension dependence of Langevin Monte Carlo. In *International Conference on Learning Representations*, 2022.

A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, volume 139, pages 8162–8171. PMLR, 2021.

A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

A. Saumard and J. A. Wellner. Log-concavity and strong log-concavity: a review. *Statistics Surveys*, 8:45–114, 2014.

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, volume 37, pages 2256–2265. PMLR, 2015.

Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Y. Song and S. Ermon. Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems*, volume 33, pages 12438–12448, 2020.

Y. Song, S. Garg, J. Shi, and S. Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, volume 124, pages 574–584. PMLR, 2020.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

W. Tang and H. Zhao. Contractive diffusion probabilistic models. *arXiv preprint arXiv:2401.13115*, 2024.

C. Villani. *Optimal Transport: Old and New*. Springer, Berlin, 2009.

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

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

Appendix A. Additional Technical Proofs

A.1 Proof of Lemma 9

Proof First of all, by following the proof of Proposition 7, we have

$$p_{T-t}(\mathbf{x}) = \int_{\mathbb{R}^d} q_1(\mathbf{x}-\mathbf{x}_0)q_0(\mathbf{x}_0)d\mathbf{x}_0, \tag{A.1}$$

where

$$q_1(\mathbf{x}) := \frac{\exp\left(-\frac{\|\mathbf{x}\|^2}{2\int_0^{T-t} e^{-2\int_s^{T-t} f(v)dv}(g(s))^2 ds}\right)}{\left(2\pi\int_0^{T-t} e^{-2\int_s^{T-t} f(v)dv}(g(s))^2 ds\right)^{d/2}},\ q_0(\mathbf{x}) := \left(e^{\int_0^{T-t} f(s)ds}\right)^d p_0\left(e^{\int_0^{T-t} f(s)ds}\mathbf{x}\right).$$

Let $\mathbf{X}_1$ and $\mathbf{X}_0$ be two independent random vectors with densities q_1 and q_0 respectively. Then it follows from (A.1) that p_{T-t} is the density of $\mathbf{X}_1 + \mathbf{X}_0$. Moreover, let us write: $q_1(\mathbf{x}) = e^{-\varphi_1(\mathbf{x})}$ and $q_0(\mathbf{x}) = e^{-\varphi_0(\mathbf{x})}$. Then it follows from the proof of Proposition 7.1. in Saumard and Wellner (2014) that

$$\begin{aligned}\nabla^2(-\log p_{T-t})(\mathbf{x}) &= -\mathrm{Var}(\nabla\varphi_0(\mathbf{X}_0)|\mathbf{X}_0+\mathbf{X}_1=\mathbf{x}) + \mathbb{E}[\nabla^2\varphi_0(\mathbf{X}_0)|\mathbf{X}_0+\mathbf{X}_1=\mathbf{x}] &\text{(A.2)}\\ &= -\mathrm{Var}(\nabla\varphi_1(\mathbf{X}_1)|\mathbf{X}_0+\mathbf{X}_1=\mathbf{x}) + \mathbb{E}[\nabla^2\varphi_1(\mathbf{X}_1)|\mathbf{X}_0+\mathbf{X}_1=\mathbf{x}]. &\text{(A.3)}\end{aligned}$$

Note that it follows from the proof of Proposition 5.1.2 that

$$\nabla^2(-\log p_{T-t})(\mathbf{x}) \succeq \left(\int_0^{T-t} e^{-2\int_s^{T-t} f(v)dv}(g(s))^2 ds + e^{-\int_0^{T-t} f(s)ds} m_0^{-1}\right)^{-1} \cdot I_d. \tag{A.4}$$

On the other hand, $\mathrm{Var}(\nabla\varphi_0(\mathbf{X}_0)|\mathbf{X}_0+\mathbf{X}_1=\mathbf{x}) \succeq 0_{d\times d}$, $\mathrm{Var}(\nabla\varphi_1(\mathbf{X}_1)|\mathbf{X}_0+\mathbf{X}_1=\mathbf{x}) \succeq 0_{d\times d}$, and $\nabla\log p_0$ is L_0-Lipschitz so that $\nabla^2\varphi_0 \preceq \left(e^{\int_0^{T-t} f(s)ds}\right)^2 L_0 \cdot I_d$, and moreover, $\nabla^2\varphi_1 \preceq \left(\int_0^{T-t} e^{-2\int_s^{T-t} f(v)dv}(g(s))^2 ds\right)^{-1} \cdot I_d$. Together with (A.2) and (A.3), we have

$$\nabla^2(-\log p_{T-t})(\mathbf{x}) \preceq \min\left(\left(\int_0^{T-t} e^{-2\int_s^{T-t} f(v)dv}(g(s))^2 ds\right)^{-1}, \left(e^{\int_0^{T-t} f(s)ds}\right)^2 L_0\right) \cdot I_d. \tag{A.5}$$

Hence, it follows from (A.4) and (A.5) that $\log p_{T-t}$ is $L(T-t)$-Lipschitz, where $L(T-t)$ is given in (3.6). This completes the proof. ■

A.2 Proof of Lemma 10

Proof We can compute that for any $(k-1)\eta \le t \le k\eta$,

$$\begin{aligned}\mathbf{z}_t - \mathbf{z}_{(k-1)\eta} &= \tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_{(k-1)\eta}\\ &\quad + \int_{(k-1)\eta}^t \left[f(T-s)(\mathbf{z}_s - \tilde{\mathbf{x}}_s) + (g(T-s))^2\left(\nabla\log p_{T-s}(\mathbf{z}_s) - \nabla\log p_{T-s}(\tilde{\mathbf{x}}_s)\right)\right] ds,\end{aligned}$$

and therefore

$$\begin{aligned}&\left\|\mathbf{z}_t-\mathbf{z}_{(k-1)\eta}\right\|_{L_2}\\&\leq\left\|\tilde{\mathbf{x}}_t-\tilde{\mathbf{x}}_{(k-1)\eta}\right\|_{L_2}+\int_{(k-1)\eta}^{t}\left[f(T-s)+(g(T-s))^2L(T-s)\right]\left\|\mathbf{z}_s-\tilde{\mathbf{x}}_s\right\|_{L_2}ds.\end{aligned}$$

We obtained in the proof of Proposition 7 that

$$\|\mathbf{z}_t-\tilde{\mathbf{x}}_t\|_{L_2}\leq e^{-\frac{1}{2}\int_0^t m(T-s)ds}\|\mathbf{z}_0-\tilde{\mathbf{x}}_0\|_{L_2},$$

and moreover, from the proof of Proposition 7, we have $\|\mathbf{z}_0-\tilde{\mathbf{x}}_0\|_{L_2}\leq e^{-\int_0^T f(s)ds}\|\mathbf{x}_0\|_{L_2}$ so that

$$\|\mathbf{z}_t-\tilde{\mathbf{x}}_t\|_{L_2}\leq e^{-\frac{1}{2}\int_0^t m(T-s)ds}e^{-\int_0^T f(s)ds}\|\mathbf{x}_0\|_{L_2}.$$

Therefore, we have

$$\left\|\mathbf{z}_t-\mathbf{z}_{(k-1)\eta}\right\|_{L_2}\leq\left\|\tilde{\mathbf{x}}_t-\tilde{\mathbf{x}}_{(k-1)\eta}\right\|_{L_2}+c_1(T)\int_{(k-1)\eta}^{k\eta}\left[f(T-s)+(g(T-s))^2L(T-s)\right]ds,\tag{A.6}$$

where $c_1(T)$ bounds $\sup_{0\leq t\leq T}\|\mathbf{z}_t-\tilde{\mathbf{x}}_t\|_{L_2}$ and it is given in (3.12). Moreover, we recall that the backward process $(\tilde{\mathbf{x}}_t)_{0\leq t\leq T}$ has the same distribution as the forward process $(\mathbf{x}_{T-t})_{0\leq t\leq T}$, so that $\left\|\tilde{\mathbf{x}}_t-\tilde{\mathbf{x}}_{(k-1)\eta}\right\|_{L_2}=\left\|\mathbf{x}_{T-t}-\mathbf{x}_{T-(k-1)\eta}\right\|_{L_2}$, where $\mathbf{x}_t$ satisfies the SDE: $d\mathbf{x}_t=-f(t)\mathbf{x}_t dt+g(t)d\mathbf{B}_t$, with $\mathbf{x}_0\sim p_0$. Therefore, we have

$$\mathbf{x}_{T-(k-1)\eta}-\mathbf{x}_{T-t}=-\int_{T-t}^{T-(k-1)\eta}f(s)\mathbf{x}_s ds+\int_{T-t}^{T-(k-1)\eta}g(s)d\mathbf{B}_s.$$

We can compute that

$$\begin{aligned}\|\mathbf{x}_{T-(k-1)\eta}-\mathbf{x}_{T-t}\|_{L_2}&\leq\int_{T-t}^{T-(k-1)\eta}f(s)\|\mathbf{x}_s\|_{L_2}ds+\left\|\int_{T-t}^{T-(k-1)\eta}g(s)d\mathbf{B}_s\right\|_{L_2}\\&\leq\sup_{0\leq t\leq T}\|\mathbf{x}_t\|_{L_2}\int_{T-k\eta}^{T-(k-1)\eta}f(s)ds+\left(\int_{T-k\eta}^{T-(k-1)\eta}(g(s))^2ds\right)^{1/2}\sqrt{d},\end{aligned}$$

where we used Itô's isometry. Therefore, we have

$$\|\mathbf{x}_{T-(k-1)\eta}-\mathbf{x}_{T-t}\|_{L_2}\leq c_2(T)\int_{T-k\eta}^{T-(k-1)\eta}f(s)ds+\left(\int_{T-k\eta}^{T-(k-1)\eta}(g(s))^2ds\right)^{1/2}\sqrt{d},$$

where we recall from (5.14) that $c_2(T)=\sup_{0\leq t\leq T}\|\mathbf{x}_t\|_{L_2}$ with an explicit formula given in (3.13). It follows that Lemma 10 holds. ■

Appendix B. Derivation of Results in Table 2

In this section, we prove the results that are summarized in Table 2. We discuss variance exploding SDEs in Appendix B.1, variance preserving SDEs in Appendix B.2, and constant coefficient SDEs in Appendix B.3.

B.1 Variance-Exploding SDEs

In this section, we consider variance-exploding SDEs with $f(t) \equiv 0$ in the forward process (1.1). We can immediately obtain the following corollary of Theorem 2.

Corollary 11 *Assume that Assumptions 1, 2, 3 and 4 hold. Then, we have*

$$\begin{aligned}\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) &\leq e^{-\int_0^{K\eta} c(t)dt}\|\mathbf{x}_0\|_{L_2} \\ &+ \sum_{k=1}^{K}\prod_{j=k+1}^{K}\gamma_{j,\eta}\cdot\left(M_1\eta\left(1+2\|\mathbf{x}_0\|_{L_2}+\sqrt{d}\left(\int_0^T (g(t))^2dt\right)^{1/2}\right)\int_{(k-1)\eta}^{k\eta}(g(T-t))^2dt\right. \\ &\left.+ M\int_{(k-1)\eta}^{k\eta}(g(T-t))^2dt + \sqrt{\eta}h_{k,\eta}\left(\int_{(k-1)\eta}^{k\eta}(g(T-t))^4(L(T-t))^2dt\right)^{1/2}\right). \quad \text{(B.1)}\end{aligned}$$

In the next few sections, we consider special functions g in Corollary 11 and derive the corresponding results in Table 2.

B.1.1 Example 1: $f(t) \equiv 0$ and $g(t) = ae^{bt}$

When $g(t) = ae^{bt}$ for some $a, b > 0$, we can obtain the following result from Corollary 11.

Corollary 12 *Let $g(t) = ae^{bt}$ for some $a, b > 0$. Then, we have $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq \mathcal{O}(\epsilon)$ after $K = \mathcal{O}\left(\frac{d\log(d/\epsilon)}{\epsilon^2}\right)$ iterations provided that $M \leq \frac{\epsilon}{\log(1/\epsilon)}$ and $\eta \leq \frac{\epsilon^2}{d}$.*

Proof Let $g(t) = ae^{bt}$ for some $a, b > 0$. First, we can compute that

$$(g(t))^2 L(t) = \min\left(\frac{(g(t))^2}{\int_0^t (g(s))^2 ds}, L_0(g(t))^2\right) = \min\left(\frac{2be^{2bt}}{e^{2bt}-1}, L_0\frac{a^2}{4b^2}(e^{2bt}-1)^2\right).$$

If $e^{2bt} \geq 2$, then $e^{2bt} - 1 \geq \frac{1}{2}e^{2bt}$ and $(g(t))^2 L(t) \leq 4b$. On the other hand, if $e^{2bt} < 2$, then $(g(t))^2 L(t) \leq L_0\frac{a^2}{4b^2}$. Therefore, for any $0 \leq t \leq T$, $(g(t))^2 L(t) \leq \max\left(4b, \frac{L_0 a^2}{4b^2}\right)$. By the definition of $c(t)$ in (3.9), we can compute that

$$c(t) = \frac{m_0(g(t))^2}{1 + m_0\int_0^t (g(s))^2 ds} = \frac{m_0 a^2 e^{2bt}}{1 + m_0\frac{a^2}{2b}(e^{2bt}-1)}. \quad \text{(B.2)}$$

This implies that

$$\int_0^t c(s)ds = \int_0^t \frac{2bm_0a^2e^{2bs}ds}{2b - m_0a^2 + m_0a^2e^{2bs}} = \log\left(\frac{2b - m_0a^2 + m_0a^2e^{2bt}}{2b}\right). \quad \text{(B.3)}$$

By letting $t = T = K\eta$ in (B.3) and using (5.19), we obtain

$$e^{-\int_0^{K\eta} c(t)dt}\|\mathbf{x}_0\|_{L_2} \leq \frac{2b}{2b - m_0a^2 + m_0a^2e^{2bK\eta}}\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right).$$

Moreover,

$$h_{k,\eta} \leq \left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right)\max\left(4b, \frac{L_0a^2}{4b^2}\right)\eta + \frac{a}{\sqrt{2b}}\left(e^{2b(T-(k-1)\eta)} - e^{2b(T-k\eta)}\right)^{1/2}\sqrt{d},$$

and for any $0 \leq t \leq T$:

$$\mu(t) \geq \frac{m_0a^2e^{2bt}}{1 + m_0\frac{a^2}{2b}(e^{2bt} - 1)} - \eta\max\left(16b^2, \frac{L_0^2a^4}{16b^4}\right) \geq M_1\eta(g(t))^2 = \eta M_1a^2e^{2bt},$$

provided that

$$\eta \leq \frac{1}{2}\frac{\min(m_0a^2, 2b)}{\max\left(16b^2, \frac{L_0^2a^4}{16b^4}\right)} + \frac{1}{2}\frac{m_0a^2}{1 + m_0\frac{a^2}{2b}(e^{2bT} - 1)}.$$

Furthermore, $\mu(t) \leq \frac{m_0a^2e^{2bt}}{1+m_0\frac{a^2}{2b}(e^{2bt}-1)} \leq \max(m_0a^2, 2b)$, so that $0 \leq \gamma_{j,\eta} \leq 1$ for every $j = 1, 2, \ldots, K$, provided that $\eta \leq \min\left(\frac{1}{\max(m_0a^2,2b)}, \frac{1}{2}\frac{\min(m_0a^2,2b)}{\max\left(16b^2, \frac{L_0^2a^4}{16b^4}\right)} + \frac{1}{2}\frac{m_0a^2}{1+m_0\frac{a^2}{2b}(e^{2bT}-1)}\right)$.

Since $1 - x \leq e^{-x}$ for any $0 \leq x \leq 1$, we conclude that

$$\begin{aligned}\prod_{j=k+1}^{K}\gamma_{j,\eta} &= \prod_{j=k+1}^{K}\left(1 - \int_{(j-1)\eta}^{j\eta}\mu(T-t)dt + M_1\eta\int_{(j-1)\eta}^{j\eta}(g(T-t))^2dt\right)\\ &\leq \prod_{j=k+1}^{K} e^{-\int_{(j-1)\eta}^{j\eta}\mu(T-t)dt + M_1\eta\int_{(j-1)\eta}^{j\eta}(g(T-t))^2dt} = e^{-\int_{k\eta}^{K\eta}\mu(T-t)dt + M_1\eta\int_{k\eta}^{K\eta}(g(T-t))^2dt}.\end{aligned} \tag{B.4}$$

Moreover,

$$\begin{aligned}\int_{k\eta}^{K\eta}\mu(T-t)dt &\geq \int_{k\eta}^{K\eta}\frac{m_0a^2e^{2b(T-t)}dt}{1 + m_0\frac{a^2}{2b}(e^{2b(T-t)} - 1)} - (K-k)\eta^2\max\left(16b^2, \frac{L_0^2a^4}{16b^4}\right)\\ &= \log\left(\frac{2b - m_0a^2 + m_0a^2e^{2b(T-k\eta)}}{2b - m_0a^2 + m_0a^2e^{2b(T-K\eta)}}\right) - (K-k)\eta^2\max\left(16b^2, \frac{L_0^2a^4}{16b^4}\right),\end{aligned}$$

and $M_1\eta\int_{k\eta}^{K\eta}(g(T-t))^2dt = M_1\eta\frac{a^2}{2b}\left(e^{2b(K-k)\eta}-1\right)$. By applying Corollary 11 with $T = K\eta$, we conclude that

$$\begin{aligned}\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K),p_0) \leq{}& \frac{2b\left(\sqrt{2d/m_0}+\|\mathbf{x}_*\|\right)}{2b-m_0a^2+m_0a^2e^{2bK\eta}} + \sum_{k=1}^{K}\frac{2be^{(K-k)\eta^2\max\left(16b^2,\frac{L_0^2a^4}{16b^4}\right)+M_1\eta\frac{a^2}{2b}(e^{2b(K-k)\eta}-1)}}{2b-m_0a^2+m_0a^2e^{2b(K-k)\eta}}\\ &\cdot\left(\left(M+M_1\eta\left(1+2\left(\sqrt{2d/m_0}+\|\mathbf{x}_*\|\right)+\sqrt{d}\frac{a}{\sqrt{2b}}(e^{2bK\eta}-1)^{1/2}\right)\right)\right.\\ &\cdot\frac{a^2}{2b}\left(e^{2b(K-(k-1))\eta)}-e^{2b(K-k)\eta)}\right)\\ &+\eta\max\left(4b,\frac{L_0a^2}{4b^2}\right)\cdot\left(\left(\sqrt{2d/m_0}+\|\mathbf{x}_*\|\right)\max\left(4b,\frac{L_0a^2}{4b^2}\right)\eta\right.\\ &\left.\left.+\frac{a}{\sqrt{2b}}\left(e^{2b(K-(k-1))\eta)}-e^{2b(K-k)\eta}\right)^{1/2}\sqrt{d}\right)\right).\end{aligned}$$

By the mean-value theorem, we have $e^{2b(K-(k-1))\eta)}-e^{2b(K-k)\eta)} \leq 2be^{2b(K-(k-1))\eta}\eta$, which implies that

$$\begin{aligned}\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K),p_0) \leq{}& \mathcal{O}\left(\frac{\sqrt{d}}{e^{2bK\eta}}\right)+\mathcal{O}\left(e^{K\eta^2\max\left(16b^2,\frac{L_0^2a^4}{16b^4}\right)+M_1\eta\frac{a^2}{2b}e^{2bK\eta}}\right.\\ &\cdot\sum_{k=1}^{K}\frac{1}{e^{2b(K-k)\eta}}\cdot\left.\left(\left(M+M_1\eta\sqrt{d}e^{bK\eta}\right)e^{2b(K-(k-1))\eta}\eta+\eta e^{b(K-(k-1))\eta}\sqrt{\eta}\sqrt{d}\right)\right)\\ \leq{}& \mathcal{O}\left(\frac{\sqrt{d}}{e^{2bK\eta}}\right)+\mathcal{O}\left(e^{K\eta^2\max\left(16b^2,\frac{L_0^2a^4}{16b^4}\right)}\cdot\left(\left(M+M_1\eta\sqrt{d}e^{bK\eta}\right)K\eta+\sqrt{\eta}\sqrt{d}\right)\right)\leq\mathcal{O}(\epsilon),\end{aligned}$$

provided that $K\eta = \frac{\log(\sqrt{d}/\epsilon)}{2b}$, $M \leq \frac{\epsilon}{\log(1/\epsilon)}$, and $\eta \leq \frac{\epsilon^2}{d}$, which implies that $K \geq \mathcal{O}\left(\frac{d\log(d/\epsilon)}{\epsilon^2}\right)$. This completes the proof. ■

Remark 13 *In Corollary 12, we can also spell out the dependence of iteration complexity on M_1 from Assumption 2. It follows from the proof of Corollary 12 that $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K),p_0)\leq\mathcal{O}(\epsilon)$ provided that $K\eta = \frac{\log(\sqrt{d}/\epsilon)}{2b}$, $M\leq\frac{\epsilon}{\log(1/\epsilon)}$, and $\eta\leq\frac{\epsilon^2}{d}$, with the additional constraint that $\eta\leq\frac{M}{M_1\sqrt{d}e^{bK\eta}}\leq\frac{\epsilon^{3/2}}{M_1\log(1/\epsilon)d^{3/4}}$, which implies that $K\geq\mathcal{O}\left(\log\left(\frac{d}{\epsilon}\right)\max\left\{\frac{d}{\epsilon^2},\frac{M_1\log(1/\epsilon)d^{3/4}}{\epsilon^{3/2}}\right\}\right)$.*

B.1.2 Example 2: $f(t)\equiv 0$ and $g(t)\equiv a$

When $g(t)\equiv a$ for some $a>0$, we can obtain the following result from Corollary 11.

Corollary 14 *Let $g(t) \equiv a$ for some $a > 0$. Then, we have $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq \mathcal{O}(\epsilon)$ after $K = \mathcal{O}\left(\frac{d^{3/2}\log(d/\epsilon)}{\epsilon^3}\right)$ iterations provided that $M \leq \mathcal{O}\left(\frac{\epsilon}{\sqrt{\log(d/\epsilon)}}\right)$ and $\eta \leq \mathcal{O}\left(\frac{\epsilon^2}{d\log(d/\epsilon)}\right)$.*

Proof When $g(t) \equiv a$, by applying Corollary 11 with $T = K\eta$ and (5.19), we have

$$\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq \frac{\sqrt{2d/m_0} + \|\mathbf{x}_*\|}{1 + m_0 a^2 K\eta} + \sum_{k=1}^{K} \frac{1}{1 + m_0 a^2 (K-k)\eta} e^{(K-k)\eta^2 L_0^2 a^4 + M_1 \eta (K-k)\eta a^2}$$
$$\cdot \Bigg(\left(M + M_1\eta\left(1 + 2\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right) + \sqrt{d}a\sqrt{K\eta}\right)\right) a^2\eta$$
$$+ \eta L_0 a^2 \cdot \left(\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right) L_0 a^2 \eta + a\sqrt{\eta}\sqrt{d}\right)\Bigg).$$

It is easy to verify that

$$\sum_{k=1}^{K} \frac{e^{(K-k)\eta^2 L_0^2 a^4 + M_1 \eta^2 (K-k) a^2}}{1 + m_0 a^2 (K-k)\eta} \leq \left(1 + \frac{\log((K-1)\eta m_0 a^2)}{\eta m_0 a^2}\right) e^{K\eta^2 L_0^2 a^4 + M_1 \eta^2 K a^2}.$$

Therefore,

$$\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq \mathcal{O}\Bigg(\frac{\sqrt{d}}{1 + m_0 a^2 K\eta} + \left(1 + \frac{\log((K-1)\eta m_0 a^2)}{\eta m_0 a^2}\right) e^{K\eta^2 L_0^2 a^4 + M_1 \eta^2 K a^2}$$
$$\cdot \left(\left(M + M_1\eta\sqrt{d}\sqrt{K\eta}\right) a^2\eta + \eta L_0 a^2 \cdot \left(\sqrt{d} L_0 a^2 \eta + a\sqrt{\eta}\sqrt{d}\right)\right)\Bigg).$$

Hence, we conclude that $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq \mathcal{O}(\epsilon)$ provided that $K\eta = \frac{\sqrt{d}}{\epsilon}$, $M = \mathcal{O}\left(\sqrt{\eta d}\right)$, and $\eta \leq \mathcal{O}\left(\frac{\epsilon^2}{d\log(d/\epsilon)}\right)$, so that $M = \mathcal{O}(\sqrt{\eta d}) \leq \mathcal{O}\left(\frac{\epsilon}{\sqrt{\log(d/\epsilon)}}\right)$, which implies that $K \geq \mathcal{O}\left(\frac{d^{3/2}\log(d/\epsilon)}{\epsilon^3}\right)$. This completes the proof. ■

B.1.3 Example 3: $f(t) \equiv 0$ and $g(t) = \sqrt{2at}$

When $g(t) = \sqrt{2at}$ for some $a > 0$, we can obtain the following result from Corollary 11.

Corollary 15 *Let $g(t) = \sqrt{2at}$ for some $a > 0$. Then, we have $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq \mathcal{O}(\epsilon)$ after $K = \mathcal{O}\left(\frac{d^{5/4}}{\epsilon^{5/2}}\right)$ iterations provided that $M \leq \epsilon^{3/2}$ and $\eta \leq \frac{\epsilon^2}{d}$.*

Proof When $g(t) = \sqrt{2at}$ for some $a > 0$, we have

$$(g(t))^2 L(t) = \min\left(\frac{(g(t))^2}{\int_0^t (g(s))^2 ds}, L_0 (g(t))^2\right) = \min\left(\frac{2}{t}, 2tL_0 a\right) \leq 2\sqrt{aL_0}.$$

We can also compute from (3.9) that

$$\int_0^t c(s)ds = \int_0^t \frac{2m_0 a s ds}{1 + m_0 a s^2} = \log\left(1 + m_0 a t^2\right). \tag{B.5}$$

By letting $t = T = K\eta$ in (B.5) and using (5.19), we obtain

$$e^{-\int_0^{K\eta} c(t)dt} \|\mathbf{x}_0\|_{L_2} \leq \frac{\sqrt{2d/m_0} + \|\mathbf{x}_*\|}{1 + a m_0 K^2 \eta^2}.$$

Moreover,

$$h_{k,\eta} \leq 2\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right)\sqrt{aL_0}\eta + \left((T - (k-1)\eta)^2 - (T - k\eta)^2\right)^{1/2}\sqrt{d},$$

and for any $0 \leq t \leq T$:

$$\mu(t) = \frac{2am_0 t}{1 + am_0 t^2} - 4\eta \min\left(\frac{1}{t^2}, t^2 a^2 L_0^2\right) \geq M_1 \eta (g(t))^2 = 2aM_1 \eta t,$$

provided that $\eta \leq \min\left(\frac{m_0}{4L_0(a^2L_0^2 + am_0)}, \frac{m_0}{4a^2L_0^2(L_0 + m_0)}, \frac{m_0}{2M_1(1 + am_0T^2)}\right)$. Additionally,

$$\mu(t) \leq \frac{2am_0 t}{1 + am_0 t^2} \leq \sqrt{am_0},$$

so that $0 \leq \gamma_{j,\eta} \leq 1$ for any $j = 1, 2, \ldots, K$, provided that

$$\eta \leq \min\left(\frac{m_0}{4L_0(a^2L_0^2 + am_0)}, \frac{m_0}{4a^2L_0^2(L_0 + m_0)}, \frac{m_0}{2M_1(1 + am_0T^2)}, \frac{1}{\sqrt{am_0}}\right).$$

We recall from (B.4) that

$$\prod_{j=k+1}^{K} \gamma_{j,\eta} \leq e^{-\int_{k\eta}^{K\eta} \mu(T-t)dt + M_1\eta \int_{k\eta}^{K\eta} (g(T-t))^2 dt}.$$

Moreover, one can verify that

$$\int_{k\eta}^{K\eta} \mu(T-t)dt \geq \log\left(1 + am_0(K-k)^2\eta^2\right) - 4(K-k)\eta^2 aL_0,$$

and

$$M_1\eta \int_{k\eta}^{K\eta} (g(T-t))^2 dt = M_1\eta \int_{k\eta}^{K\eta} 2a(K\eta - t)dt = M_1\eta(K-k)^2\eta^2.$$

By applying Corollary 11 with $T = K\eta$ and (5.19), we conclude that

$$\begin{aligned}
\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq{}& \frac{\sqrt{2d/m_0} + \|\mathbf{x}_*\|}{1 + am_0K^2\eta^2} + \sum_{k=1}^{K} \frac{1}{1 + am_0(K-k)^2\eta^2} e^{4K\eta^2 aL_0 + M_1\eta^3K^2} \\
&\cdot \Bigg(2\left(M + M_1\eta\left(1 + 2\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right) + \sqrt{d}\sqrt{2a}K\eta\right)\right)(K-k+1)\eta^2 \\
&+ 2\sqrt{aL_0}\eta\left(2\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right)\sqrt{aL_0}\eta + (2(K-k+1))^{1/2}\eta\sqrt{d}\right)\Bigg).
\end{aligned}$$

This implies that

$$\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq \mathcal{O}\left(\frac{\sqrt{d}}{K^2\eta^2} + e^{\mathcal{O}((K\eta)^2\eta)}\left(K\eta(M + M_1\sqrt{d}K\eta^2) + \sqrt{\eta}\sqrt{d}\right)\right) \leq \mathcal{O}(\epsilon),$$

provided that $K\eta = \frac{d^{1/4}}{\sqrt{\epsilon}}$, $M \leq \epsilon^{3/2}$, and $\eta \leq \frac{\epsilon^2}{d}$, so that $K \geq \mathcal{O}\left(\frac{d^{5/4}}{\epsilon^{5/2}}\right)$. This completes the proof. ■

B.1.4 Example 4: $f(t) \equiv 0$ and $g(t) = (b + at)^c$

When $g(t) = (b+at)^c$ for some $a, b, c > 0$, we obtain the following result from Corollary 11.

Corollary 16 *Let $g(t) = (b + at)^c$ for some $a, b, c > 0$. Then, we have $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq \mathcal{O}(\epsilon)$ after $K = \mathcal{O}\left(\frac{d^{\frac{1}{2(2c+1)}+1}}{\epsilon^{\frac{1}{2c+1}+2}}\right)$ iterations provided that $M \leq \epsilon^{1+\frac{2c}{2c+1}}$ and $\eta \leq \frac{\epsilon^2}{d}$.*

Proof When $g(t) = (b + at)^c$ for some $a, b, c > 0$, we can compute that

$$(g(t))^2 L(t) = \min\left(\frac{(b+at)^c}{\frac{1}{a(2c+1)}((b+at)^{2c+1} - b^{2c+1})}, L_0(b+at)^{2c}\right). \tag{B.6}$$

It is straightforward to verify that $(g(t))^2 L(t) \leq \max\left(\frac{a(2c+1)}{(1-\frac{1}{2^{2c+1}})b}, L_0(2b)^{2c}\right)$. By (3.9), we have

$$e^{-\int_0^{K\eta} c(t)dt}\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right) = \frac{\sqrt{2d/m_0} + \|\mathbf{x}_*\|}{1 + \frac{m_0}{a(2c+1)}((b+aK\eta)^{2c+1} - b^{2c+1})}.$$

Also,

$$h_{k,\eta} \leq \left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right)\max\left(\frac{a(2c+1)}{(1-\frac{1}{2^{2c+1}})b}, L_0(2b)^{2c}\right)\eta + \left(\frac{(b+a(T-(k-1)\eta))^{2c}}{a}\right)^{1/2}\sqrt{\eta}\sqrt{d},$$

and for any $0 \leq t \leq T$:

$$\mu(t) = \frac{(b+at)^{2c}}{\frac{1}{m_0} + \frac{1}{a(2c+1)}((b+at)^{2c+1} - b^{2c+1})} - \eta\min\left(\frac{(b+at)^{2c}}{\frac{1}{a^2(2c+1)^2}((b+at)^{2c+1} - b^{2c+1})^2}, L_0^2(b+at)^{4c}\right) \geq M_1\eta(g(t))^2 = M_1\eta(b+at)^{2c},$$

provided that $\eta \le \frac{1}{2}$, $\eta \le \frac{\frac{b^{2c}}{\frac{1}{m_0}+\frac{1}{\sqrt{m_0}}+1}}{2L_0^2(a(2c+1)(\frac{1}{\sqrt{m_0}}+1)+b^{2c+1})^{\frac{4c}{2c+1}}}$ and $\eta \le \frac{1}{\frac{2}{m_0}+\frac{2}{a(2c+1)}((b+aT)^{2c+1}-b^{2c+1})}$. In addition,

$$\mu(t) \le \frac{(b+at)^{2c}}{\frac{1}{m_0}+\frac{1}{a(2c+1)}((b+at)^{2c+1}-b^{2c+1})} \le \max\left(\frac{a(2c+1)}{b}, m_0 b^{2c}\right),$$

so that $0 \le \gamma_{j,\eta} \le 1$ for any $j = 1, 2, \ldots, K$. We recall from (B.4) that

$$\prod_{j=k+1}^{K} \gamma_{j,\eta} \le e^{-\int_{k\eta}^{K\eta} \mu(T-t)dt + M_1\eta \int_{k\eta}^{K\eta} (g(T-t))^2 dt}.$$

Moreover,

$$\int_{k\eta}^{K\eta} \mu(T-t)dt \ge \log\left(1 + \frac{m_0}{a(2c+1)}\left((b+a(K-k)\eta)^{2c+1} - b^{2c+1}\right)\right) - (K-k)\eta^2 \max\left(\frac{a^2(2c+1)^2}{(1-\frac{1}{2^{2c+1}})^2 b^2}, L_0^2(2b)^{4c}\right),$$

and we can compute that

$$M_1\eta \int_{k\eta}^{K\eta} (g(T-t))^2 dt = \frac{M_1\eta}{a(2c+1)}\left((b+a(K-k)\eta)^{2c+1} - b^{2c+1}\right).$$

By applying Corollary 11 with $T = K\eta$ and (5.19), we conclude that

$$\begin{aligned}
\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \le{}& \frac{\sqrt{2d/m_0} + \|\mathbf{x}_*\|}{1 + \frac{m_0}{a(2c+1)}((b+aK\eta)^{2c+1} - b^{2c+1})} \\
&+ \sum_{k=1}^{K} \frac{e^{\frac{(K-k)\eta^2 \max\left(\frac{a^2(2c+1)^2}{(1-\frac{1}{2^{2c+1}})^2 b^2}, L_0^2(2b)^{4c}\right) + \frac{M_1\eta}{a(2c+1)}((b+a(K-k)\eta)^{2c+1} - b^{2c+1})}}{1 + \frac{m_0}{a(2c+1)}\left((b+a(K-k)\eta)^{2c+1} - b^{2c+1}\right)} \\
&\cdot \left(\left(M + M_1\eta\left(1 + 2\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right) + \sqrt{d}\left(\frac{(b+aK\eta)^{2c+1} - b^{2c+1}}{a(2c+1)}\right)^{1/2}\right)\right)\right. \\
&\qquad \cdot \frac{(b+a(K-k+1)\eta)^{2c+1} - (b+a(K-k)\eta)^{2c+1}}{a(2c+1)} \\
&\qquad + \eta \max\left(\frac{a(2c+1)}{(1-\frac{1}{2^{2c+1}})b}, L_0(2b)^{2c}\right) \\
&\qquad \cdot \left(\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right) \max\left(\frac{a(2c+1)}{(1-\frac{1}{2^{2c+1}})b}, L_0(2b)^{2c}\right)\eta \right. \\
&\qquad\qquad \left.\left. + \left(\frac{(b+a((K-(k-1))\eta))^{2c}}{a}\right)^{1/2} \sqrt{\eta}\sqrt{d}\right)\right).
\end{aligned}$$

This implies that

$$\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq \mathcal{O}\left(\frac{\sqrt{d}}{(K\eta)^{2c+1}} + e^{\mathcal{O}((K\eta)\eta + (K\eta)^{2c+1}\eta)}\left((K\eta)^{2c}M + \sqrt{\eta}\sqrt{d}\right)\right) \leq \mathcal{O}(\epsilon),$$

provided that $K\eta = \frac{d^{\frac{1}{2(2c+1)}}}{\epsilon^{\frac{1}{2c+1}}}$, $M \leq \epsilon^{1+\frac{2c}{2c+1}}$, and $\eta \leq \frac{\epsilon^2}{d}$, so that $K \geq \mathcal{O}\left(\frac{d^{\frac{1}{2(2c+1)}+1}}{\epsilon^{\frac{1}{2c+1}+2}}\right)$. This completes the proof. ■

B.2 Variance-Preserving SDEs

In this section, we consider Variance-Preserving SDEs with $f(t) = \frac{1}{2}\beta(t)$ and $g(t) = \sqrt{\beta(t)}$ in the forward process (1.1). We can obtain the following corollary of Theorem 2.

Corollary 17 *Under the assumptions of Theorem 2, we have*

$$\begin{aligned}
\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq{}& \frac{\|\mathbf{x}_0\|_{L_2}}{m_0 e^{\int_0^{K\eta}\beta(s)ds} + 1 - m_0} \\
&+ \sum_{k=1}^{K} \frac{e^{\int_{k\eta}^{K\eta}\frac{1}{2}\beta(K\eta-t)dt + \int_{k\eta}^{K\eta}\frac{\eta}{4}(\beta(K\eta-t))^2dt + \int_{k\eta}^{K\eta}4\eta\max(1,L_0^2)(\beta(K\eta-t))^2dt + M_1\eta\int_{k\eta}^{K\eta}\beta(K\eta-t)dt}}{m_0 e^{\int_0^{(K-k)\eta}\beta(s)ds} + 1 - m_0} \\
&\quad \cdot \Bigg(M_1\eta(1 + 2\|\mathbf{x}_0\|_{L_2} + \sqrt{d}) \int_{(k-1)\eta}^{k\eta} \beta(K\eta - t)dt \\
&+ M\int_{(k-1)\eta}^{k\eta}\beta(K\eta-t)dt + \sqrt{\eta}\left(\frac{1}{2} + 2\max(1, L_0)\right)\left(\int_{(k-1)\eta}^{k\eta}(\beta(K\eta-t))^2dt\right)^{1/2} \\
&\quad \cdot \Bigg(e^{-\int_0^{K\eta}\frac{1}{2}\beta(s)ds}\|\mathbf{x}_0\|_{L_2}\left(\frac{1}{2} + 2\max(1, L_0)\right)\int_{(k-1)\eta}^{k\eta}\beta(K\eta - s)ds \\
&\quad + \left(\|\mathbf{x}_0\|_{L_2}^2 + d\right)^{1/2}\int_{(K-k)\eta}^{(K-(k-1))\eta}\frac{1}{2}\beta(s)ds + \left(\int_{(K-k)\eta}^{(K-(k-1))\eta}\beta(s)ds\right)^{1/2}\sqrt{d}\Bigg)\Bigg). \quad \text{(B.7)}
\end{aligned}$$

Proof We apply Theorem 2 applied to the variance-preserving SDE ($f(t) = \frac{1}{2}\beta(t)$ and $g(t) = \sqrt{\beta(t)}$). First, we can compute that

$$L(T-t) = \min\left(\frac{1}{1 - e^{-\int_0^{T-t}\beta(s)ds}}, e^{\int_0^{T-t}\beta(s)ds}L_0\right).$$

If $e^{\int_0^{T-t}\beta(s)ds} \geq 2$, then $\frac{1}{1-e^{-\int_0^{T-t}\beta(s)ds}} \leq 2$ and otherwise $e^{\int_0^{T-t}\beta(s)ds}L_0 \leq 2L_0$. Therefore, for any $0 \leq t \leq T$, $L(T-t) \leq 2\max(1, L_0)$. By applying Theorem 2, we have

$$\begin{aligned}
\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) &\leq e^{-\int_0^{K\eta} c(t)dt}\|\mathbf{x}_0\|_{L_2} + \sum_{k=1}^{K}\prod_{j=k+1}^{K}\gamma_{j,\eta} \\
&\cdot \Bigg(M_1\eta\left(1 + \|\mathbf{x}_0\|_{L_2} + c_2(T)\right)\int_{(k-1)\eta}^{k\eta}(g(T-t))^2dt \\
&+ M\int_{(k-1)\eta}^{k\eta}(g(T-t))^2dt + \sqrt{\eta}h_{k,\eta}\left(\int_{(k-1)\eta}^{k\eta}[f(T-t) + (g(T-t))^2L(T-t)]^2dt\right)^{1/2}\Bigg),
\end{aligned}$$

where, the definition of $\gamma_{j,\eta}$ in (3.14) depends on $\mu(T-t)$, such that for any $0 \leq t \leq T$:

$$\begin{aligned}
\mu(T-t) &= \frac{(g(T-t))^2}{\frac{1}{m_0}e^{-2\int_0^{T-t}f(s)ds} + \int_0^{T-t}e^{-2\int_s^{T-t}f(v)dv}(g(s))^2ds} - f(T-t) \\
&\qquad - \eta(f(T-t))^2 - \eta(g(T-t))^4(L(T-t))^2, \\
&\geq \frac{m_0\beta(T-t)}{e^{-\int_0^{T-t}\beta(s)ds} + m_0(1 - e^{-\int_0^{T-t}\beta(s)ds})} \\
&\qquad - \frac{1}{2}\beta(T-t) - \frac{\eta}{4}(\beta(T-t))^2 - 4\eta(\beta(T-t))^2\max(1, L_0^2),
\end{aligned}$$

where we assume η is sufficiently small such that $0 \leq \gamma_{j,\eta} \leq 1$, for every $j = 1, 2, \ldots, K$. One can verify that

$$\begin{aligned}
h_{k,\eta} &\leq e^{-\int_0^T \frac{1}{2}\beta(s)ds}\|\mathbf{x}_0\|_{L_2}\left(\frac{1}{2} + 2\max(1, L_0)\right)\int_{(k-1)\eta}^{k\eta}\beta(T-s)ds \\
&\quad + \left(\|\mathbf{x}_0\|_{L_2}^2 + d\right)^{1/2}\int_{T-k\eta}^{T-(k-1)\eta}\frac{1}{2}\beta(s)ds + \left(\int_{T-k\eta}^{T-(k-1)\eta}\beta(s)ds\right)^{1/2}\sqrt{d}.
\end{aligned}$$

Next, for VP-SDE, we have $f(t) = \frac{1}{2}\beta(t)$ and $g(t) = \sqrt{\beta(t)}$ so that we can compute:

$$c(t) = \frac{m_0\beta(t)}{e^{-\int_0^t\beta(s)ds} + m_0\int_0^t e^{-\int_s^t\beta(v)dv}\beta(s)ds} = \frac{m_0\beta(t)}{e^{-\int_0^t\beta(s)ds} + m_0(1 - e^{-\int_0^t\beta(s)ds})}.$$

It follows that

$$\int_0^T c(t)dt = \int_0^{\int_0^T\beta(s)ds}\frac{m_0dx}{m_0 + (1-m_0)e^{-x}} = \log\left(m_0e^{\int_0^T\beta(s)ds} + 1 - m_0\right).$$

Hence, we obtain

$$e^{-\int_0^T c(t)dt}\|\mathbf{x}_0\|_{L_2} = \frac{\|\mathbf{x}_0\|_{L_2}}{m_0e^{\int_0^T\beta(s)ds} + 1 - m_0}.$$

By using $T = K\eta$, we complete the proof. ■

Next, we prove Proposition 4.

B.2.1 Proof of Proof of Proposition 4

Proof It follows from Corollary 17 and (5.19) that

$$\begin{aligned}
\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq & \frac{\sqrt{2d/m_0} + \|\mathbf{x}_*\|}{m_0 e^{\int_0^{K\eta} \beta(s)ds} + 1 - m_0} \\
& + \sum_{k=1}^{K} \frac{e^{(1+2M_1\eta)\int_0^{(K-k)\eta} \frac{1}{2}\beta(t)dt + (\frac{\eta}{4} + 4\eta\max(1, L_0^2))\int_0^{(K-k)\eta} (\beta(t))^2 dt}}{m_0 e^{\int_0^{(K-k)\eta} \beta(s)ds} + 1 - m_0} \\
& \cdot \left(\left(M + M_1\eta \left(1 + 2\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\| \right) + \sqrt{d} \right) \right) \int_{(K-k)\eta}^{(K-k+1)\eta} \beta(t)dt \right. \\
& + \sqrt{\eta} \left(\frac{1}{2} + 2\max(1, L_0) \right) \left(\int_{(K-k)\eta}^{(K-k+1)\eta} (\beta(t))^2 dt \right)^{1/2} \\
& \cdot \left(e^{-\int_0^{K\eta} \frac{1}{2}\beta(s)ds} \left(\sqrt{2d/m_0} + \|\mathbf{x}_*\| \right) \left(\frac{1}{2} + 2\max(1, L_0) \right) \int_{(K-k)\eta}^{(K-k+1)\eta} \beta(s)ds \right. \\
& \left. \left. + \left(\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\| \right)^2 + d \right)^{1/2} \int_{(K-k)\eta}^{(K-(k-1))\eta} \frac{1}{2}\beta(s)ds + \left(\int_{(K-k)\eta}^{(K-(k-1))\eta} \beta(s)ds \right)^{1/2} \sqrt{d} \right) \right).
\end{aligned}$$

Since $\beta(t)$ is increasing in t, we can compute

$$\begin{aligned}
\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq \mathcal{O}\Bigg(& \frac{\sqrt{d}}{e^{\int_0^{K\eta} \beta(s)ds}} + e^{M_1\eta \int_0^{K\eta} \beta(t)dt + (\frac{\eta}{4} + 4\eta\max(1, L_0^2))\beta(K\eta) \int_0^{K\eta} \beta(t)dt} \\
& \cdot \left(\left(M + M_1\eta\sqrt{d} \right) \beta(K\eta) + \beta(K\eta) \cdot \left(\sqrt{d}\eta\beta(K\eta) + \sqrt{\beta(K\eta)}\sqrt{\eta}\sqrt{d} \right) \right) \Bigg).
\end{aligned}$$

Since $\beta(t) \leq c_1 \left(\int_0^t \beta(s)ds \right)^{c_3} + c_2$ uniformly in t for some $c_1, c_2, c_3 > 0$, we have

$$\begin{aligned}
\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq \mathcal{O}\Bigg(& \frac{\sqrt{d}}{e^{\int_0^{K\eta} \beta(s)ds}} + e^{M_1\eta \int_0^{K\eta} \beta(t)dt + (\frac{\eta}{4} + 4\eta\max(1, L_0^2)) \left(c_1 \left(\int_0^{K\eta} \beta(t)dt \right)^{1+c_3} + c_2 \int_0^{K\eta} \beta(t)dt \right)} \\
& \cdot \left(\left(M + M_1\eta\sqrt{d} \right) \left(\int_0^{K\eta} \beta(t)dt \right)^{c_3} \right. \\
& \left. \left. + \left(\sqrt{d}\eta \left(\int_0^{K\eta} \beta(t)dt \right)^{2c_3} + \left(\int_0^{K\eta} \beta(t)dt \right)^{3c_3/2} \sqrt{\eta}\sqrt{d} \right) \right) \right) \leq \mathcal{O}(\epsilon),
\end{aligned}$$

provided that $\int_0^{K\eta} \beta(s)ds = \log\left(\sqrt{d}/\epsilon\right)$, $M \leq \frac{\epsilon}{(\log(\sqrt{d}/\epsilon))^{c_3}}$, and $\eta \leq \frac{\epsilon^2}{d(\log(1/\epsilon))^{3/c_3}}$. Since $\beta(t)$ is increasing, $\log\left(\sqrt{d}/\epsilon\right) \geq \beta(0)K\eta$, so that $K \leq \frac{\log(d/\epsilon)}{\beta(0)\eta} = \mathcal{O}\left(\frac{d(\log(d/\epsilon))^{3c_3+1}}{\epsilon^2} \right)$ if we take $\eta = \frac{\epsilon^2}{d(\log(d/\epsilon))^{3/c_3}}$. This completes the proof. ■

In the next two subsections, we consider special functions $\beta(t)$ in Corollary 17 and derive the corresponding results in Table 2.

B.2.2 Example 1: $\beta(t) = (b + at)^\rho$

We first consider the special case $\beta(t) = (b+at)^\rho$. This includes the special case $\beta(t) = b+at$ when $\rho = 1$ that is studied in Ho et al. (2020).

Corollary 18 *Assume $\beta(t) = (b + at)^\rho$. Then, we have $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq \mathcal{O}(\epsilon)$ after $K = \mathcal{O}\left(\frac{d(\log(d/\epsilon))^{\frac{1}{\rho+1}}}{\epsilon^2}\right)$ iterations provided that $M \leq \epsilon$ and $\eta \leq \frac{\epsilon^2}{d}$.*

Proof When $\beta(t) = (b + at)^\rho$, by Corollary 17 and (5.19), we can compute that

$$
\begin{aligned}
\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq{}& \frac{\sqrt{2d/m_0} + \|\mathbf{x}_*\|}{m_0 e^{\frac{1}{a(\rho+1)}((b+aK\eta)^{\rho+1} - b^{\rho+1})} + 1 - m_0} \\
&+ \sum_{k=1}^{K} \frac{e^{\frac{1+2M_1\eta}{2a(\rho+1)}((b+a(K-k)\eta)^{\rho+1} - b^{\rho+1}) + (\frac{\eta}{4} + 4\eta\max(1, L_0^2))\frac{1}{a(2\rho+1)}((b+a(K-k)\eta)^{2\rho+1} - b^{2\rho+1})}}{m_0 e^{\frac{1}{a(\rho+1)}((b+a(K-k)\eta\eta)^{\rho+1} - b^{\rho+1})} + 1 - m_0} \\
&\cdot \Bigg(\left(M + M_1\eta\left(1 + 2\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right) + \sqrt{d}\right)\right) \\
&\qquad \cdot \frac{\left((b + a(K-k+1)\eta)^{\rho+1} - (b + a(K-k)\eta)^{\rho+1}\right)}{a(\rho+1)} \\
&+ \sqrt{\eta}\left(\frac{1}{2} + 2\max(1, L_0)\right)\left(\frac{\left((b + a(K-k+1)\eta)^{2\rho+1} - (b + a(K-k)\eta)^{2\rho+1}\right)}{a(2\rho+1)}\right)^{1/2} \\
&\cdot \Bigg(e^{-\frac{((b+aK\eta)^{\rho+1} - b^{\rho+1})}{2a(\rho+1)}}\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right)\left(\frac{1}{2} + 2\max(1, L_0)\right) \\
&\qquad \cdot \frac{\left((b + a(K-k+1)\eta)^{\rho+1} - (b + a(K-k)\eta)^{\rho+1}\right)}{a(\rho+1)} \\
&+ \frac{1}{2}\Bigg(\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right)^2 + d\Bigg)^{1/2} \frac{\left((b + a(K-k+1)\eta)^{\rho+1} - (b + a(K-k)\eta)^{\rho+1}\right)}{a(\rho+1)} \\
&\qquad + \left(\frac{1}{a(\rho+1)}\left((b + a(K-k+1)\eta)^{\rho+1} - (b + a(K-k)\eta)^{\rho+1}\right)\right)^{1/2}\sqrt{d}\Bigg)\Bigg).
\end{aligned}
$$

Thus, we can compute that

$$
\begin{aligned}
\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq{}& \mathcal{O}\Bigg(\frac{\sqrt{d}}{e^{\frac{(b+aK\eta)^{\rho+1}}{a(\rho+1)}}} + \sum_{k=1}^{K} \frac{e^{\frac{1+2M_1\eta}{2a(\rho+1)}((b+a(K-k)\eta)^{\rho+1} - b^{\rho+1}) + (\frac{\eta}{4} + 4\eta\max(1, L_0^2))\frac{(b+aK\eta)^{2\rho+1}}{a(2\rho+1)}}}{m_0 e^{\frac{1}{a(\rho+1)}((b+a(K-k)\eta\eta)^{\rho+1} - b^{\rho+1})} + 1 - m_0} \\
&\cdot \Bigg(\left(M + M_1\eta\left(1 + 2\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right) + \sqrt{d}\right)\right)((K-k+1)\eta)^\rho\eta \\
&+ \sqrt{\eta}((K-k+1)\eta)^\rho\sqrt{\eta} \cdot \left(\sqrt{d}((K-k+1)\eta)^\rho\eta + ((K-k+1)\eta)^{\rho/2}\sqrt{\eta}\sqrt{d}\right)\Bigg)\Bigg) \\
\leq{}& \mathcal{O}\Bigg(\frac{\sqrt{d}}{e^{\frac{(b+aK\eta)^{\rho+1}}{a(\rho+1)}}} + e^{(\frac{\eta}{4} + 4\eta\max(1, L_0^2))\frac{(b+aK\eta)^{2\rho+1}}{a(2\rho+1)}} \cdot \left(M + M_1\eta\sqrt{d} + \left(\sqrt{d}\eta + \sqrt{\eta}\sqrt{d}\right)\right)\Bigg) \leq \mathcal{O}(\epsilon),
\end{aligned}
$$

provided that $K\eta = \frac{(a(\rho+1))^{\frac{1}{\rho+1}}}{a}\left(\log\left(\sqrt{d}/\epsilon\right)\right)^{\frac{1}{\rho+1}} - \frac{b}{a}$, $M \leq \epsilon$, and $\eta \leq \frac{\epsilon^2}{d}$, which implies that $K \geq \mathcal{O}\left(\frac{d(\log(d/\epsilon))^{\frac{1}{\rho+1}}}{\epsilon^2}\right)$. This completes the proof. ■

Remark 19 *In Corollary 18, we can also spell out the dependence of iteration complexity on M_1 from Assumption 2. It follows from the proof of Corollary 18 that $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq \mathcal{O}(\epsilon)$ provided that $K\eta = \frac{(a(\rho+1))^{\frac{1}{\rho+1}}}{a}\left(\log\left(\sqrt{d}/\epsilon\right)\right)^{\frac{1}{\rho+1}} - \frac{b}{a}$, $M \leq \epsilon$, and $\eta \leq \frac{\epsilon^2}{d}$, with the additional constraint that $\eta \leq \frac{M}{M_1\sqrt{d}} \leq \frac{\epsilon}{M_1\sqrt{d}}$, which implies that $K \geq \mathcal{O}\left(\left(\log\left(\frac{d}{\epsilon}\right)\right)^{\frac{1}{\rho+1}} \max\left\{\frac{d}{\epsilon^2}, \frac{M_1\sqrt{d}}{\epsilon}\right\}\right)$.*

B.2.3 Example 2: $\beta(t) = ae^{bt}$

Corollary 20 *Assume $\beta(t) = ae^{bt}$. Then, we have $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq \mathcal{O}(\epsilon)$ after $K = \mathcal{O}\left(\frac{d\log(\log(d/\epsilon))}{\epsilon^2}\right)$ iterations provided that $M \leq \epsilon$ and $\eta \leq \frac{\epsilon^2}{d}$.*

Proof When $\beta(t) = ae^{bt}$, by Corollary 17 and (5.19), we can compute that

$$
\begin{aligned}
\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq\ & \frac{\sqrt{2d/m_0} + \|\mathbf{x}_*\|}{m_0 e^{\frac{a}{b}(e^{bK\eta}-1)} + 1 - m_0} \\
& + \sum_{k=1}^{K} \frac{e^{(1+2M_1\eta)\frac{a}{2b}(e^{b(K-k)\eta}-1) + (\frac{\eta}{4} + 4\eta\max(1,L_0^2))\frac{a^2}{2b}(e^{2b(K-k)\eta}-1)}}{m_0 e^{\frac{a}{b}(e^{b(K-k)\eta}-1)} + 1 - m_0} \\
& \quad \cdot \Bigg(\left(M + M_1\eta\left(1 + 2\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right) + \sqrt{d}\right)\right)\frac{a}{b}\left(e^{b(K-k+1)\eta} - e^{b(K-k)\eta}\right) \\
& \quad + \sqrt{\eta}\left(\frac{1}{2} + 2\max(1, L_0)\right)\left(\frac{a^2}{2b}\left(e^{2b(K-k+1)\eta} - e^{2b(K-k)\eta}\right)\right)^{1/2} \\
& \cdot \Bigg(e^{-\frac{a}{2b}(e^{bK\eta}-1)}\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right)\left(\frac{1}{2} + 2\max(1, L_0)\right)\frac{a}{b}\left(e^{b(K-k+1)\eta} - e^{b(K-k)\eta}\right) \\
& + \frac{1}{2}\Bigg(\left(\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right)^2 + d\right)^{1/2}\frac{a}{b}\left(e^{b(K-k+1)\eta} - e^{b(K-k)\eta}\right) \\
& \qquad + \left(\frac{a}{b}\left(e^{b(K-k+1)\eta} - e^{b(K-k)\eta}\right)\right)^{1/2}\sqrt{d}\Bigg)\Bigg).
\end{aligned}
$$

Thus, we can compute that

$$\begin{aligned}\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq \mathcal{O}\Bigg(& \frac{\sqrt{d}}{e^{\frac{a}{b}e^{bK\eta}}} + \sum_{k=1}^{K} \frac{e^{(1+2M_1\eta)\frac{a}{2b}(e^{b(K-k)\eta}-1)+(\frac{\eta}{4}+4\eta\max(1,L_0^2))\frac{a^2}{2b}(e^{2b(K-k)\eta}-1)}}{m_0 e^{\frac{a}{b}(e^{b(K-k)\eta}-1)} + 1 - m_0} \\ & \cdot \Bigg(\left(M + M_1\eta\left(1 + 2\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right) + \sqrt{d}\right)\right) e^{b(K-k)\eta}\eta \\ & \quad + \sqrt{\eta}e^{b(K-k)\eta}\sqrt{\eta}\left(e^{-\frac{a}{2b}e^{bK\eta}}e^{b(K-k)\eta}\eta + \sqrt{d}e^{b(K-k)\eta}\eta + e^{\frac{1}{2}b(K-k)\eta}\sqrt{\eta}\sqrt{d}\right)\Bigg)\Bigg) \\ \leq \mathcal{O}\Bigg(& \frac{\sqrt{d}}{e^{\frac{a}{b}e^{bK\eta}}} + e^{(\frac{\eta}{4}+4\eta\max(1,L_0^2))\frac{a^2}{2b}e^{2bK\eta}} \cdot \left(M + M_1\eta\sqrt{d} + \left(\sqrt{d}\eta + \sqrt{\eta}\sqrt{d}\right)\right)\Bigg) \leq \mathcal{O}(\epsilon),\end{aligned}$$

provided that $K\eta = \frac{1}{b}\log\left(\frac{b}{a}\log\left(\frac{\sqrt{d}}{\epsilon}\right)\right)$, $M \leq \epsilon$, and $\eta \leq \frac{\epsilon^2}{d}$, which implies that $K \geq \mathcal{O}\left(\frac{d\log(\log(d/\epsilon))}{\epsilon^2}\right)$. This completes the proof. ■

B.3 Constant Coefficient SDE

Corollary 21 *Under the assumptions of Theorem 2, assume that $f(t) \equiv \alpha > 0$ and $g(t) \equiv \sigma > 0$. Further assume that $m_0 \geq \frac{2\alpha}{\sigma^2}$ and $\eta \leq \min\left\{1, \frac{\alpha}{2\alpha^2+2(2\alpha+\sigma^2L_0)^2}\right\}$. Then, we have*

$$\begin{aligned}\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq & \frac{\sqrt{2d/m_0} + \|\mathbf{x}_*\|}{\frac{m_0\sigma^2(e^{2\alpha K\eta}-1)}{2\alpha} + 1} + \frac{2}{\alpha}M_1\eta\left(1 + 2\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right) + \frac{\sigma\sqrt{d}}{\sqrt{2\alpha}}\right)\sigma^2 \\ & + \frac{2}{\alpha}\left(M\sigma^2 + \eta^{1/2}\tilde{C}_1\left(3\alpha + \sigma^2L_0\right)\right),\end{aligned}$$

where $\tilde{C}_1 := \left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right)(4\alpha + \sigma^2L_0) + \alpha\sqrt{\frac{d\sigma^2}{2\alpha}} + \sigma\sqrt{d}$. In particular, for any given $\epsilon > 0$, we have $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq \epsilon$ provided that $M \leq \frac{\epsilon\alpha}{8\sigma^2}$,

$$\eta \leq \min\left(\frac{\epsilon^2\alpha^2}{64\tilde{C}_1^2\left(3\alpha + \sigma^2L_0\right)^2}, \frac{\alpha}{8M_1\left(1 + 2\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right) + \frac{\sigma\sqrt{d}}{\sqrt{2\alpha}}\right)\sigma^2}\right),$$

and $K\eta \geq \frac{1}{2\alpha}\log\left(\left(\frac{4\left(\sqrt{2d/m_0}+\|\mathbf{x}_*\|\right)}{\epsilon} - 1\right)\frac{2\alpha}{m_0\sigma^2} + 1\right)$.

Remark 22 *Corollary 21 implies that $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \leq \epsilon$ by choosing (if we just keep track of the dependence on ϵ and d) $M = \mathcal{O}(\epsilon)$, $\eta = \mathcal{O}\left(\frac{\epsilon^2}{d}\right)$, and $K \geq \mathcal{O}\left(\frac{d}{\epsilon^2}\log\left(\frac{d}{\epsilon}\right)\right)$.*

B.3.1 Proof of Corollary 21

Proof Using a similar argument as in previous sections, one can readily obtain the following upper bound on $h_{k,\eta}$ in (3.15) for the special case $f(t) \equiv \alpha > 0$ and $g(t) \equiv \sigma > 0$:

$$\begin{aligned} h_{k,\eta} &\le \eta \mathcal{W}_2(p_0, \hat{p}_T)\left(3\alpha + \sigma^2 L_0\right) + \eta\alpha\left(\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right)^2 + \frac{d\sigma^2}{2\alpha}\right)^{1/2} + \sqrt{\eta}\sigma\sqrt{d} \\ &\le \eta\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right)\left(3\alpha + \sigma^2 L_0\right) + \eta\alpha\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right) + \eta\alpha\sqrt{\frac{d\sigma^2}{2\alpha}} + \sqrt{\eta}\sigma\sqrt{d} \\ &= \sqrt{\eta}C_1, \end{aligned}$$

where

$$C_1 := \sqrt{\eta}\left(\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right)\left(4\alpha + \sigma^2 L_0\right) + \alpha\sqrt{\frac{d\sigma^2}{2\alpha}}\right) + \sigma\sqrt{d}. \tag{B.8}$$

Moreover, we recall the formula for $\mu(T-t)$ from (3.10) so that we can compute that

$$\begin{aligned} &\mu(T-t) - M_1\eta(g(T-t))^2 \\ &\ge \frac{\sigma^2}{\frac{1}{m_0}e^{-2\alpha(T-t)} + \frac{\sigma^2}{2\alpha}\left(1 - e^{-2\alpha(T-t)}\right)} - \alpha - \eta\alpha^2 - \eta\sigma^4(2\alpha\sigma^{-2} + L_0)^2 - M_1\eta\sigma^2 \ge \frac{\alpha}{2}, \end{aligned}$$

provided that $m_0 \ge \frac{2\alpha}{\sigma^2}$ and $\eta \le \frac{\alpha}{2\alpha^2 + 2\sigma^4(2\alpha\sigma^{-2}+L_0)^2 + 2M_1\sigma^2}$. Since $c(t) = \frac{\sigma^2}{\frac{e^{-2\alpha t}}{m_0} + \frac{\sigma^2}{2\alpha}(1-e^{-2\alpha t})} > 0$, we have $\int_0^{K\eta} c(t)dt = \log\left(\frac{m_0\sigma^2(e^{2\alpha K\eta}-1)}{2\alpha} + 1\right)$. Hence, by Theorem 2 and (5.19), we have

$$\begin{aligned} &\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \\ &\le \frac{\sqrt{2d/m_0} + \|\mathbf{x}_*\|}{\frac{m_0\sigma^2(e^{2\alpha K\eta}-1)}{2\alpha} + 1} + \sum_{k=1}^{K}\left(1 - \frac{\alpha\eta}{2}\right)^{K-k} \cdot \left(M_1\eta\left(1 + 2\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right) + \frac{\sigma}{\sqrt{2\alpha}}\sqrt{d}\right)\eta\sigma^2\right) \\ &\quad + \sum_{k=1}^{K}\left(1 - \frac{\alpha\eta}{2}\right)^{K-k} \cdot \left(M\eta\sigma^2 + \eta^{3/2}C_1\left(\alpha + \sigma^2\left(2\alpha\sigma^{-2} + L_0\right)\right)\right) \\ &\le \frac{\sqrt{2d/m_0} + \|\mathbf{x}_*\|}{\frac{m_0\sigma^2(e^{2\alpha K\eta}-1)}{2\alpha} + 1} \\ &\quad + \frac{2}{\alpha}\left(M_1\eta\left(1 + 2\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right) + \frac{\sigma\sqrt{d}}{\sqrt{2\alpha}}\right)\sigma^2 + M\sigma^2 + \eta^{1/2}C_1\left(3\alpha + \sigma^2 L_0\right)\right), \end{aligned}$$

where

$$C_1 \le \tilde{C}_1 := \left(\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right)\left(4\alpha + \sigma^2 L_0\right) + \alpha\sqrt{\frac{d\sigma^2}{2\alpha}}\right) + \sigma\sqrt{d},$$

where C_1 is defined in (B.8) and we used the assumption that the stepsize $\eta \le 1$. In particular, given any $\epsilon > 0$, we have $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_K), p_0) \le \epsilon$ if we take $M \le \frac{\epsilon\alpha}{8\sigma^2}$, $\eta \le \frac{\epsilon^2\alpha^2}{64\tilde{C}_1^2(3\alpha+\sigma^2 L_0)^2}$, $\eta \le \frac{\alpha}{8M_1\left(1+2\left(\sqrt{2d/m_0}+\|\mathbf{x}_*\|\right)+\frac{\sigma\sqrt{d}}{\sqrt{2\alpha}}\right)\sigma^2}$, and $K\eta \ge \frac{1}{2\alpha}\log\left(\left(\frac{4\left(\sqrt{2d/m_0}+\|\mathbf{x}_*\|\right)}{\epsilon} - 1\right)\frac{2\alpha}{m_0\sigma^2} + 1\right)$. This completes the proof. ■