Convergence Analysis for General Probability Flow ODEs of Diffusion Models in Wasserstein Distances

Xuefeng Gao
Department of Systems Engineering
and Engineering Management
Chinese University of Hong Kong

Lingjiong Zhu
Department of Mathematics
Florida State University

Abstract

Score-based generative modeling with probability flow ordinary differential equations (ODEs) has achieved remarkable success in a variety of applications. While various fast ODE-based samplers have been proposed in the literature and employed in practice, the theoretical understandings about convergence properties of the probability flow ODE are still quite limited. In this paper, we provide the first non-asymptotic convergence analysis for a general class of probability flow ODE samplers in 2-Wasserstein distance, assuming accurate score estimates and smooth log-concave data distributions. We then consider various examples and establish results on the iteration complexity of the corresponding ODE-based samplers. Our proof technique relies on spelling out explicitly the contraction rate for the continuous-time ODE and analyzing the discretization and score-matching errors using synchronous coupling; the challenge in our analysis mainly arises from the inherent non-autonomy of the probability flow ODE and the specific exponential integrator that we study.

1 Introduction

Score-based generative models (SGMs) (Sohl-Dickstein et al., 2015; Song and Ermon, 2019; Ho et al., 2020; Song et al., 2021), or diffusion models, have achieved remarkable success in a range of applications, particularly in the realm of image and audio generation (Rombach

et al., 2022; Ramesh et al., 2022; Popov et al., 2021). These models employ a unique approach where samples from a target data distribution are progressively corrupted with noise through a forward process. Subsequently, the models learn to reverse this corrupted process in order to generate new samples.

In this paper, we aim to provide theoretical guarantees for the probability flow ODE (ordinary differential equation) implementation of SGMs, proposed initially in Song et al. (2021). The forward process in SGMs, denoted by $(\mathbf{x}_t)_{t\in[0,T]}$, follows the stochastic differential equation (SDE):

$$\mathrm{d}\mathbf{x}_t = -f(t)\mathbf{x}_t\mathrm{d}t + g(t)\mathrm{d}\mathbf{B}_t, \quad \mathbf{x}_0 \sim p_0, \qquad (1)$$

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 $d-$dimensional (unknown) target data distribution. Two popular choices of forward processes in the literature are Variance Exploding (VE) SDEs and Variance Preserving (VP) SDEs (Song et al., 2021); see Section 2 for more details. If we denote $p_t(\mathbf{x})$ as the probability density function of $\mathbf{x}_t$ in (1), then Song et al. (2021) showed that there exists an ODE:

$$\frac{\mathrm{d}\tilde{\mathbf{x}}_t}{\mathrm{d}t} = f(T-t)\tilde{\mathbf{x}}_t + \frac{1}{2}(g(T-t))^2\nabla_{\mathbf{x}}\log p_{T-t}(\tilde{\mathbf{x}}_t),$$
$$\tilde{\mathbf{x}}_0 \sim p_T, \qquad (2)$$

where the solution $\tilde{\mathbf{x}}_t$ at time $t \in [0,T]$, is distributed according to p_{T-t}. The ODE (2) is called the *probability flow ODE*. Note that the probability flow ODE involves the *score function*, $\nabla_{\mathbf{x}}\log p_t(\mathbf{x})$, which is unknown. In practice, it can be approximated using neural networks which are trained with appropriate score-matching techniques (Hyvärinen and Dayan, 2005; Vincent, 2011; Song et al., 2020). Once the score function is estimated, one can sample $\tilde{\mathbf{x}}_0$ from a normal distribution to initialize the ODE, and numerically solve the ODE forward in time with any ODE solvers such as Euler Song et al. (2021) or Heun's 2nd order method Karras

et al. (2022). The resulting sample generated at time T can be viewed as an approximate sample from the target distribution since $\tilde{\mathbf{x}}_T \sim p_0$.

A large body of work on diffusion models has recently investigated various methods for faster generation of samples based on the probability flow ODE; see, e.g., (Lu et al., 2022; Karras et al., 2022; Zhang and Chen, 2023; Zhao et al., 2023; Song et al., 2023). While these methods are quite effective in practice, the theoretical understandings of this probability flow ODE approach are still quite limited. To our best knowledge, Chen et al. (2023e) established the first non-asymptotic convergence guarantees for the probability flow ODE sampler in the Kullback-Leibler (KL) divergence, but it did not provide concrete polynomial dependency on the dimension d and $1/\epsilon$, where ϵ is the prescribed error between the data distribution and the generated distribution. Chen et al. (2023c) considered a specific VP-SDE as the forward process (where $f \equiv 1$ and $g \equiv \sqrt{2}$), and provided polynomial-time guarantees for some variants of the probability flow ODE in the total variation (TV) distance. Specifically, the algorithms they analyzed rely on the use of stochastic corrector steps, so the samplers are not fully deterministic. Li et al. (2024b) also considered a specific VP-SDE as the forward processes and analyzed directly a discrete-time version of probability flow ODEs and obtained convergence guarantees for fully deterministic samplers in TV. Very recently, Li et al. (2024a) extended Li et al. (2024b) and established nonasymptotic convergence guarantees in TV for the accelerated DDIM-type deterministic sampler.

These theoretical studies have mostly focused on convergence analysis of probability flow ODE samplers in TV distance between the data distribution and the generated distribution. However, practitioners are often interested in the 2-Wasserstein ($\mathcal{W}_2$) distance. For instance, in image-related tasks, Fréchet Inception Distance (FID) is a widely adopted performance metric for evaluating the quality of generated samples, where FID measures the $\mathcal{W}_2$ distance between the distribution of generated images and the distribution of real images Heusel et al. (2017). We also note that TV distance does not upper bound 2-Wasserstein distance on $\mathbb{R}^d$ (see e.g. Gibbs and Su (2002)). In addition, previous works (Chen et al., 2023c; Li et al., 2024b,a) have studied specific choices of f and g in their convergence analysis of ODE-based samplers. However, it has been shown in Song et al. (2021) that the empirical performance of probability flow ODE samplers depends crucially on the choice of f and g in (2), which indicates the importance of selecting these parameters (noise schedules) of diffusion models. This leads to the following question which we study in this paper:

Can we establish Wasserstein convergence guarantees for probability flow ODE samplers with general functions f and g?

Our Contributions.

- We establish non-asymptotic convergence guarantees for a general class of probability flow ODEs in 2-Wasserstein distance, assuming that the score function can be accurately learned and the data distribution has a smooth and strongly log-concave density (Theorem 3.4). In particular, we allow general functions f and g in the probability flow ODE (2), and our results apply to both VP and VE SDE models. Theorem 3.4 directly translates to an upper bound on the iteration complexity, which is the number of sampling steps needed for the ODE sampler to yield $\epsilon-$accuracy in 2-Wasserstein distance between the data distribution and the generative distribution of the SGMs.

- We specialize our result to ODE samplers with specific functions f and g that are commonly used in the literature, and we find the complexity of VE-SDEs is worse than that of VP-SDEs for the examples we analyze (see Table 1 for details). This theoretical finding is consistent with the empirical observation in Song et al. (2021), where they found that the empirical performance of probability flow ODE samplers depends on the choice of f and g, and the sample quality for VE-SDEs is much worse than VP-SDEs for high-dimensional data.

- We obtain an iteration complexity bound $\widetilde{\mathcal{O}}\left(\sqrt{d}/\epsilon\right)$ of the ODE sampler for each of the three examples of VP-SDEs studied, where d is the dimension of the data distribution and $\widetilde{\mathcal{O}}$ ignores the logarithmic factors and hides dependency on other parameters. We also show that (see Proposition 3.12) under mild assumptions there are no other choices of f and g so that the iteration complexity can be better than $\widetilde{\mathcal{O}}\left(\sqrt{d}/\epsilon\right)$.

- Our main proof technique relies on spelling out an explicit contraction rate in the continuous-time ODE and providing a careful analysis in controlling the discretization and score-matching errors. Our proof technique is inspired by iteration complexity results in the context of Langevin algorithms for sampling in the literature (Dalalyan and Karagulyan, 2019); the smooth strong log-concavity of data distribution allows us to obtain contraction for the probability flow ODE using the synchronous coupling. Yet our analysis is more sophisticated and subtle even under the smooth

strong log-concavity assumption. First, in the literature of sampling strongly-log-concave target densities, where Langevin algorithms are often used, the underlying dynamics is time-homogeneous, and the strong log-concavity and smoothness assumptions can be directly used (see e.g. Dalalyan and Karagulyan (2019)), whereas for probability flow ODEs, it is non-autonomous, and the strong-log-concavity and smoothness constants need to be carefully analyzed and spelled out which are time-dependent. Second, the usage of exponential integrator also makes the analysis more subtle, as compared to the Euler discretization of the ODE. We incorporate the exponential integrator and the interpolation in the discretization to design an continuous-time auxiliary ODE, and then control the error between this auxiliary ODE and the probability flow ODE. Finally, in order to further analyze the VE-SDE and VP-SDE examples in Section 3.3 (see Corollaries 3.8-3.11, Proposition 3.12), the inherent non-autonomous property of the probability flow ODE makes the analysis much trickier and subtle. One has to perform a series of inequalities based on the formula in Theorem 3.4 in order to spell out the leading order terms to obtain the iteration complexities.

1.1 Related Work

In addition to the deterministic sampler based on probability flow ODEs, another major family of diffusion samplers is based on discretization of the reverse-time SDE, which is obtained by reversing the forward process (1) in time (Anderson, 1982; Cattiaux et al., 2023)). This leads to SDE-based stochastic samplers due to the noise in the reverse-time SDE. Compared with SDE-based samplers, ODE-based deterministic samplers are often claimed to converge much faster, at the cost of slightly inferior sample quality; see e.g. Yang et al. (2023). In recent years, there has been a significant surge in research focused on the convergence theory of SDE-based stochastic samplers for diffusion models, particularly when assuming access to precise estimates of the score function; see, e.g., Block et al. (2020); De Bortoli et al. (2021); De Bortoli (2022); Lee et al. (2022, 2023); Chen et al. (2023a,d); Li et al. (2024b); Chen et al. (2023b); Gao et al. (2025); Benton et al. (2024a); Tang and Zhao (2024). These studies have mostly focused on the convergence analysis of SDE-based stochastic samplers in TV or KL divergence. (De Bortoli, 2022; Chen et al., 2023a,d) provided Wasserstein convergence bound for the SDE-based sampler for the DDPM model in Ho et al. (2020) for bounded data distribution, in which case the 2-Wasserstein distance can be bounded by the TV distance. Gao et al. (2025) established convergence guarantees for SDE-based samplers for a general class of

SGMs in 2-Wasserstein distance, for unbounded smooth log-concave data distribution. Our study differs from these studies in that we consider deterministic samplers based on the probability flow ODE implementation of SGMs.

Our work is also broadly related to flow based methods or flow matching, see e.g. Lipman et al. (2022); Albergo and Vanden-Eijnden (2022). The flow matching framework is more general than the probability flow ODE, because it approximates a flow between two arbitrary distributions. In particular, flow matching reduces to probability flow ODE for diffusion models when the starting distribution is Gaussian. The theoretical analysis of flow matching methods are still limited, and there are currently very few error bounds in Wasserstein distance for flow matching methods with a fully deterministic sampling scheme. Two very recent studies on such bounds are Benton et al. (2024b) and Albergo et al. (2023). While both studies have considered the errors due to the approximate flow/velocity, they do not consider the error arising from numerically solving the ODE. Hence, these results are not about the iteration complexity of the sampling scheme, which we study for the probability flow ODE. Cheng et al. (2023) recently provided convergence guarantees for a progressive flow model, which differs from score-based diffusion models in that the flow model they consider is deterministic in the forward (data-to-noise) process.

Notations. For any d-dimensional random vector $\mathbf{x}$ with finite second moment, the L_2-norm of $\mathbf{x}$ is defined as $\|\mathbf{x}\|_{L_2} = \sqrt{\mathbb{E}\|\mathbf{x}\|^2}$, where $\|\cdot\|$ denotes the Euclidean norm. We denote $\mathcal{L}(\mathbf{x})$ as the law of $\mathbf{x}$. For any two Borel probability measures $\mu_1, \mu_2 \in \mathcal{P}_2(\mathbb{R}^d)$, the space consisting of all the Borel probability measures on $\mathbb{R}^d$ with the finite second moment (based on the Euclidean norm), the standard 2-Wasserstein distance Villani (2009) is defined by

$$\mathcal{W}_2(\mu_1, \mu_2) := \sqrt{\inf \mathbb{E}\left[\|\mathbf{x}_1 - \mathbf{x}_2\|^2\right]},$$

where the infimum is taken over all joint distributions of the random vectors $\mathbf{x}_1, \mathbf{x}_2$ with marginal distributions μ_1, μ_2. A differentiable function F from $\mathbb{R}^d$ to $\mathbb{R}$ is said to be μ-strongly convex and L-smooth (i.e. ∇F is L-Lipschitz) if for every $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$,

$$\frac{\mu}{2}\|\mathbf{u} - \mathbf{v}\|^2 \le F(\mathbf{u}) - F(\mathbf{v}) - \nabla F(\mathbf{v})^\top (\mathbf{u} - \mathbf{v})$$

$$\le \frac{L}{2}\|\mathbf{u} - \mathbf{v}\|^2.$$

2 Preliminaries

Recall $p_0 \in \mathcal{P}(\mathbb{R}^d)$ denotes the unknown data distribution which has a density, 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 problem of generative modelling is to generate new samples that (approximately) follow the data distribution.

We consider the probability flow ODE (see (2)) based implementation of SGMs for sample generation, see e.g. Song et al. (2021). The functions f and g in (2) can be general in our study, and in particular, our study covers the following two classes of models commonly used in the literature: (1) $f(t) \equiv 0$ 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. The corresponding forward SDE (1) is referred to as Variance Exploding (VE) SDE. (2) $f(t) = \frac{1}{2}\beta(t)$ and $g(t) = \sqrt{\beta(t)}$ for some nondecreasing function $\beta(t)$, e.g., $\beta(t) = at+b$ for some positive constants a, b. The corresponding forward SDE (1) is referred to as Variance Preserving (VP) SDE.

To implement the ODE (2), one needs to (a) sample from a tractable distribution (aka prior distribution) to initialize the ODE, (b) estimate the score function $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$, and (c) numerically solve/integrate the ODE. We next discuss these three issues (or error sources).

First, we discuss (a), the prior distribution. Note that p_T is unknown. To provide an unifying analysis for probability flow ODEs with general f and g (including both VE and VP SDE models), we choose the prior distribution to be a normal distribution $\hat{p}_T$ given as follows:

$$\hat{p}_T := \mathcal{N}\left(0, \int_0^T e^{-2\int_t^T f(s)ds}(g(t))^2 dt \cdot I_d\right), \quad (3)$$

and I_d is the d-dimensional identity matrix. To see why this is a reasonable choice, we note that the forward SDE (1) has an explicit solution

$$\mathbf{x}_t = e^{-\int_0^t f(s)ds}\mathbf{x}_0 + \int_0^t e^{-\int_s^t f(v)dv}g(s)d\mathbf{B}_s. \quad (4)$$

Hence, we can the take the distribution of the Brownian integral $\int_0^t e^{-\int_s^t f(v)dv}g(s)d\mathbf{B}_s$ in (4), which is $\hat{p}_T$, as an approximation of p_T, so that

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

(See Lemma B.3 in the Appendix for the details.) Hence, we consider the ODE:

$$\frac{d\mathbf{y}_t}{dt} = f(T-t)\mathbf{y}_t + \frac{1}{2}(g(T-t))^2\nabla \log p_{T-t}(\mathbf{y}_t),$$
$$\mathbf{y}_0 \sim \hat{p}_T, \quad (5)$$

as an approximation of the ODE (2) which starts from p_T.

Remark 2.1. If the forward process is a VP-SDE with a stationary distribution which is normal, one can also take it as the prior distribution and our main result in this paper can be adapted to this setting.

We next consider (b) score matching, i.e., approximate the unknown true score function $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ using a time-dependent score model $s_\theta(\mathbf{x}, t)$, which is often a deep neural network parameterized by θ. To train the score model, one can use, for instance, denoising score matching Song et al. (2021), where the training objective for optimizing the neural network is given by

$$\min_\theta \int_0^T \Big[\lambda(t)\mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{\mathbf{x}_t|\mathbf{x}_0}\Big\|s_\theta(\mathbf{x}_t, t)$$
$$- \nabla_{\mathbf{x}_t} \log p_{t|0}(\mathbf{x}_t|\mathbf{x}_0)\Big\|^2\Big]dt. \quad (6)$$

Here, $\lambda(\cdot) : [0, T] \to \mathbb{R}_{>0}$ is some positive weighting function, $\mathbf{x}_0 \sim p_0$ is the data distribution, and $p_{t|0}(\mathbf{x}_t|\mathbf{x}_0)$ is the density of $\mathbf{x}_t$ given $\mathbf{x}_0$, which is Gaussian due to the choice of the forward SDE in (1). Because one has i.i.d. samples from p_0, the distribution of $\mathbf{x}_0$, the objective in (6) can be approximated by Monte Carlo methods and the resulting loss function can be then optimized.

After the score function is estimated, we introduce a continuous-time process that approximates (5):

$$\frac{d\mathbf{z}_t}{dt} = f(T-t)\mathbf{z}_t + \frac{1}{2}(g(T-t))^2 s_\theta(\mathbf{z}_t, T-t),$$
$$\mathbf{z}_0 \sim \hat{p}_T, \quad (7)$$

where we replace the true score function in (5) by the estimated score s_θ.

Finally, we discuss (c) numerically solve the ODE (7) for generation of new samples. There are various methods proposed and employed in practice, including Euler (Song et al., 2021), Heun's 2nd order method (Karras et al., 2022), DPM solver (Lu et al., 2022), exponential integrator (Zhang and Chen, 2023), to name just a few. In this paper, we consider the following exponential integrator (i.e. exactly integrating the linear part) discretization of the ODE (7) for our theoretical convergence analysis. Let $\eta > 0$ be the stepsize and without loss of generality, let us assume that $T = K\eta$, where K is a positive integer. Next, we introduce an exponential integrator discretization of the ODE (7). By freezing the nonlinear nonlinear term (i.e. the $s_\theta(\mathbf{z}_t, T-t)$ term on RHS of (7)) and letting the linear term flow (i.e. the $\mathbf{z}_t$ term on RHS of (7)), we obtain the following ODE approximation of ODE (7): for any $(k-1)\eta \le t < k\eta$,

$$\frac{d\hat{\mathbf{z}}_t}{dt} = f(T-t)\hat{\mathbf{z}}_t + \frac{1}{2}(g(T-t))^2 s_\theta(\hat{\mathbf{z}}_{(k-1)\eta}, T-(k-1)\eta).$$

By solving the above (linear) ODE for $(k-1)\eta \le t < k\eta$,

we have

$$
\hat{\mathbf{z}}_{k\eta} = e^{\int_{(k-1)\eta}^{k\eta} f(T-t)\mathrm{d}t} \hat{\mathbf{z}}_{(k-1)\eta}
$$
$$
+ \frac{1}{2} \boldsymbol{s}_\theta(\hat{\mathbf{z}}_{(k-1)\eta}, T-(k-1)\eta)
$$
$$
\cdot \int_{(k-1)\eta}^{k\eta} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \mathrm{d}t. \quad (8)
$$

By letting $\mathbf{u}_k = \hat{\mathbf{z}}_{k\eta}$ for any k, we obtain the iterations for the exponential integrator discretization $(\mathbf{u}_k)_{k=0}^\infty$ of the ODE (7): for any $k = 1, 2, \ldots, K$,

$$
\mathbf{u}_k = e^{\int_{(k-1)\eta}^{k\eta} f(T-t)\mathrm{d}t} \mathbf{u}_{k-1}
$$
$$
+ \frac{1}{2} \boldsymbol{s}_\theta(\mathbf{u}_{k-1}, T-(k-1)\eta)
$$
$$
\cdot \int_{(k-1)\eta}^{k\eta} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \mathrm{d}t, \quad (9)
$$

where $\mathbf{u}_0 \sim \hat{p}_T$.

We are interested in the convergence of the generated distribution $\mathcal{L}(\mathbf{u}_K)$ to the data distribution p_0, where $\mathcal{L}(\mathbf{u}_K)$ denotes the law or distribution of $\mathbf{u}_K$. Specifically, our goal is to bound the 2-Wasserstein distance $\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0)$, and investigate the number of iterates K that is needed in order to achieve ϵ accuracy, i.e. $\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0) \leq \epsilon$.

3 Main Results

In this section we state our main results. The proofs are deferred to the Appendix.

3.1 Assumptions

The first assumption is on the density of data distribution p_0, which implies that $\mathbf{x}_0 \sim p_0$ is L_2-integrable.

Assumption 3.1. Assume that the density p_0 is differentiable and positive everywhere. Moreover, $-\nabla \log p_0$ is m_0-strongly convex and L_0-smooth for some $m_0, L_0 > 0$.

The assumption of strong-log-concavity data distribution has also been imposed in (Bruno et al., 2023; Gao et al., 2025) for convergence analysis of diffusion models. We need this assumption mainly because we consider Wasserstein convergence analysis of ODE-based deterministic samplers. In particular, the ODE (5) may not have a contraction in 2-Wasserstein distance without such an assumption; see Remark 3.7 in Section 3.2 for details.

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 the score function $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is Lipschtiz in time, where the

Lipschitz constant has at most linear growth in $\|\mathbf{x}\|$. Assumption 3.2 is needed in controlling the discretization error of the ODE (7). For Gaussian data distributions p_0, one can compute the score function $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ analytically based on (4) and readily verify that this assumption holds.

Assumption 3.2. There exists some constant L_1 such that for all $\mathbf{x}$:

$$
\sup_{\substack{1 \leq k \leq K \\ (k-1)\eta \leq t \leq k\eta}} \left\| \nabla \log p_{T-t}(\mathbf{x}) - \nabla \log p_{T-(k-1)\eta}(\mathbf{x}) \right\|
$$
$$
\leq L_1 \eta (1 + \|\mathbf{x}\|). \quad (10)
$$

Most studies on convergence analysis of diffusion models assume some form of L_2 error for score learning and focus on the sampling phase. Our next assumption is on this score-matching error. Recall $(\mathbf{u}_k)$ are the iterates defined in (9).

Assumption 3.3. Assume that there exists $M > 0$ such that

$$
\sup_{k=1,\ldots,K} \left\| \nabla \log p_{T-(k-1)\eta}(\mathbf{u}_{k-1}) \right.
$$
$$
\left. - \boldsymbol{s}_\theta(\mathbf{u}_{k-1}, T-(k-1)\eta) \right\|_{L_2} \leq M. \quad (11)
$$

Assumption 3.3 is related to the score perturbation lemma in the seminal work (Chen et al., 2023c). However, they need to assume the Hessian of score $\nabla^2 \log p_t(x)$ is bounded by L for any t and x, where L is independent of T, to obtain the desired dependency on dimension d and ϵ.

3.2 Main Result

We are now ready to state our main result.

Theorem 3.4. *Suppose that Assumptions 3.1, 3.2 and 3.3 hold and the stepsize $\eta \leq \bar{\eta}$, where $\bar{\eta} > 0$ has an explicit formula given in (27) in Appendix A. Then,*

$$
\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0) \leq \underbrace{e^{-\int_0^{K\eta} \mu(t)\mathrm{d}t} \cdot \|\mathbf{x}_0\|_{L_2}}_{\text{Initialization error}}
$$
$$
+ \underbrace{E_1(f, g, K, \eta, L_1)}_{\text{Discretization error}} + \underbrace{E_2(f, g, K, \eta, M, L_1)}_{\text{Score matching error}}.
$$
$$
(12)
$$

Here, $\mu(t)$ is given in (24) in Appendix A and

$$
E_1(f, g, K, \eta, L_1)
$$
$$
:= \sum_{k=1}^{K} \prod_{j=k+1}^{K} \gamma_{j,\eta} \cdot e^{\int_{k\eta}^{K\eta} f(T-t)\mathrm{d}t}
$$
$$
\cdot \left(\frac{L_1}{2}\eta \left(1 + \|\mathbf{x}_0\|_{L_2} + \omega(T)\right) \phi_{k,\eta} + \frac{\sqrt{\eta}}{2}\nu_{k,\eta}\sqrt{\psi_{k,\eta}} \right),
$$
(13)

$$
E_2(f, g, K, \eta, M, L_1)
$$
$$
:= \sum_{k=1}^{K} \prod_{j=k+1}^{K} \gamma_{j,\eta} \cdot e^{\int_{k\eta}^{K\eta} f(T-t)\mathrm{d}t} \cdot \frac{M}{2}\phi_{k,\eta},
$$
(14)

where $\phi_{k,\eta}$ is given in (21), $\psi_{k,\eta}$ is given in (22), $\gamma_{j,\eta}$ is given in (23), $L(t)$ is given in (26), $\delta_j(T-t)$ is defined in (30), $\omega(T)$ is defined in (32) and $\nu_{k,\eta}$ is given in (33) in Appendix A.

While the bound in Theorem 3.4 looks quite complex, it can be easily interpreted as follows.

The first term in (12), referred to as the initialization error, characterizes the convergence of the continuous-time probability flow ODE ($\mathbf{y}_t$) in (5) to the distribution p_0 without discretization or score-matching errors. Specifically, it bounds the error $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_T), p_0)$, which is introduced due to the initialization of the probability flow ODE ($\mathbf{y}_t$) at $\hat{p}_T$ instead of p_T (see Proposition 4.1). One can find from the definition (24) that $\mu(t) > 0$ and hence the initialization error goes to zero when we pick $T = K\eta$ to be sufficiently large, i.e., $e^{-\int_0^{K\eta} \mu(t)\mathrm{d}t} \cdot \|\mathbf{x}_0\|_{L_2} \to 0$, as $K\eta \to \infty$.

The second and third terms in (12) quantify the discretization and score-matching errors in running the algorithm ($\mathbf{u}_k$) in (9). Note that the assumption $\eta \leq \bar{\eta}$ in Theorem 3.4 implies that $\gamma_{j,\eta} \in (0,1)$ in (23), which plays the role of a contraction rate of the error $\|\mathbf{y}_{k\eta} - \mathbf{u}_k\|_{L_2}$ over iterations (see Proposition 4.2). Conceptually, it guarantees that as the number of iterations increases, the discretization and score-matching errors in the iterates ($\mathbf{u}_k$) do not propagate and grow in time. One can show that for fixed $T = K\eta$, the discretization error $E_1(f, g, K, \eta, L_1) \to 0$, when $\eta \to 0$ and the score matching error $E_2(f, g, K, \eta, M, L_1)$ is linear in M which goes to 0 as $M \to 0$.

By combing the above three terms, we infer that we can first choose a large $T = K\eta$, and then choose a small stepsize η so that $\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0)$ can be made small (provided that M is small).

Remark 3.5. Theorem 3.4 holds for the ODE-based sampler with exponential integrator discretization. Our analysis also goes through for the simple Euler method. For other methods such as Heun's 2nd order solver

and DPM solver, the methodology used in the current paper cannot be directly applied. In particular, we need to study their discretization errors and it will require different analysis.

Remark 3.6 (Comparison of iteration complexities). Chen et al. (2023c) analyzed probability flow ODE with stochastic corrector steps in TV and established iteration complexity of $\tilde{O}(d/\epsilon^2)$ (respectively $\tilde{O}(\sqrt{d}/\epsilon)$) when the stochastic corrector step is based on the overdamped (respectively underdamped) Langevin diffusion. Li et al. (2024b) analyzed a discrete time version of fully deterministic ODE and established an iteration complexity of $\tilde{O}(d^2/\epsilon + d^3/\sqrt{\epsilon})$ in TV. Both studies consider specific VP-SDEs as forward processes. Our paper consider fully deterministic ODE-based samplers and Wasserstein convergence guarantees. Note that TV plus strong log-concavity does not imply 2-Wasserstein convergence. For VP-SDEs, Theorem 3.4 implies that the iteration complexity is $\tilde{O}(\sqrt{d}/\epsilon)$ (for the examples considered) under our assumptions; see Table 1 for details.

Remark 3.7. Assumption 3.1 plays two roles for obtaining the upper bound in Theorem 3.4.

First, the m_0-strong-convexity of $-\nabla \log p_0$ guarantees that $\mu(t) > 0$ (which appears in the first term of (12)), which guarantees the 2-Wasserstein contraction and the convergence of the continuous-time probability flow ODE ($\mathbf{y}_t$) in (5) to the distribution p_0 without discretization or score-matching errors. Indeed, one can easily verify from (24) that $\mu(t) \to 0$ as $m_0 \to 0$, which indicates that strong-convexity of $-\nabla \log p_0$ is necessary; otherwise the ODE (5) will not have a contraction. Chen et al. (2023c) addressed this issue by modifying the ODE sampler and adding stochastic correcter steps via Langevin dynamics to establish TV convergence under weaker assumptions on data distributions. The samplers they study hence are not fully deterministic as in standard probability flow ODEs. By contrast, we analyzed the fully deterministic ODE sampler and established Wasserstein convergence.

Second, the m_0-strong-convexity of $-\nabla \log p_0$, together with the L_0-Lipschitzness of $\nabla \log p_0$, guarantees that the discretization and score-matching error at each iterate k can be explicitly controlled as in Theorem 3.4. In particular, Assumption 3.1 guarantees that $\gamma_{j,\eta} \in (0,1)$ when the stepsize $\eta \leq \bar{\eta}$, which controls the propagation of the discretization and score-matching errors as the number of iterates grows. The m_0-strong-convexity of $-\nabla \log p_0$ is necessary since one can easily verify that $\bar{\eta} \to 0$ and $\gamma_{j,\eta} \notin (0,1)$ as $m_0 \to 0$.

3.3 Examples

In this section, we consider several examples of the forward processes and discuss the implications of Theorem 3.4. In particular, we consider a variety of choices for f and g in the forward SDE (1), and investigate the iteration complexity, i.e., the number of iterates K that is needed in order to achieve ϵ accuracy, i.e. $\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0) \leq \epsilon$. While the bound in Theorem 3.4 is quite sophisticated in general, it can be made more explicit when we consider special f and g.

Example 1. We first consider a VE-SDE example from Song et al. (2021). Let $f(t) \equiv 0$ and $g(t) = ae^{bt}$ for some $a, b > 0$, we can obtain the following corollary from Theorem 3.4.

Corollary 3.8. *Let $f(t) \equiv 0$ and $g(t) = ae^{bt}$ for some $a, b > 0$. Then, we have $\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0) \leq \mathcal{O}(\epsilon)$ after $K = \mathcal{O}\left(\frac{d^{3/2}\log(d/\epsilon)}{\epsilon^3}\right)$ iterations provided that $M \leq \frac{\epsilon^2}{\sqrt{d}}$ and $\eta \leq \frac{\epsilon^3}{d^{3/2}}$.*

Example 2. Next, we consider another VE-SDE example from Gao et al. (2025), where $f \equiv 0$ and g has polynomial growth in time. This example is inspired by Karras et al. (2022) that considers $f(t) \equiv 0, g(t) = \sqrt{2t}$, where the discretization time steps are defined according to a polynomial noise schedule.

Corollary 3.9. *Let $f(t) \equiv 0$ and $g(t) = (b + at)^c$ for some $a, b > 0$, $c \geq 1/2$. Then, we have $\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0) \leq \mathcal{O}(\epsilon)$ after $K = \mathcal{O}\left(\frac{d^{\frac{1}{(2c+1)}+\frac{3}{2}}}{\epsilon^{\frac{2}{2c+1}+3}}\right)$ iterations provided that $M \leq \frac{\epsilon^2}{\sqrt{d}}$ and $\eta \leq \frac{\epsilon^3}{d^{\frac{3}{2}}}$.*

Example 3. Next, we consider a VP-SDE example, with constant f, g. This includes the special case $f \equiv 1$, $g \equiv \sqrt{2}$ that is considered in Chen et al. (2023c). In particular, we consider $f \equiv \frac{b}{2}, g \equiv \sqrt{b}$ for some $b > 0$. We obtain the following corollary from Theorem 3.4.

Corollary 3.10. *Assume $f \equiv \frac{b}{2}, g \equiv \sqrt{b}$ for some $b > 0$. Then, we have $\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0) \leq \mathcal{O}(\epsilon)$ after $K = \mathcal{O}\left(\frac{\sqrt{d}}{\epsilon}(\log(\frac{d}{\epsilon}))^2\right)$ iterations provided that $M \leq \frac{\epsilon}{\log(\sqrt{d}/\epsilon)}$ and $\eta \leq \frac{\epsilon}{\sqrt{d}\log(\sqrt{d}/\epsilon)}$.*

Example 4. Finally, we consider a VP-SDE example where f, g have polynomial growth. We consider $f(t) = \frac{1}{2}(b+at)^\rho$ and $g(t) = \sqrt{(b + at)^\rho}$, where $a, b > 0$, which is proposed in Gao et al. (2025). This includes the special case $f(t) = \frac{1}{2}(b + at)$ and $g(t) = \sqrt{b + at}$, i.e. $\rho = 1$, that is studied in Ho et al. (2020). Then we can obtain the following corollary from Theorem 3.4.

Corollary 3.11. *Assume $f(t) = \frac{1}{2}(b+at)^\rho$ and $g(t) = \sqrt{(b+at)^\rho}$. Then, we have $\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0) \leq \mathcal{O}(\epsilon)$ after $K = \mathcal{O}\left(\frac{\sqrt{d}}{\epsilon}(\log(\frac{d}{\epsilon}))^{\frac{\rho+2}{\rho+1}}\right)$ iterations provided that $M \leq \frac{\epsilon}{\log(\sqrt{d}/\epsilon)}$ and $\eta \leq \frac{\epsilon}{\sqrt{d}\log(\sqrt{d}/\epsilon)}$.*

We observe from Corollary 3.11 that the complexity K deceases as ρ increases. However, this does not suggest that the optimal complexity is achieved when $\rho \to \infty$ since our complexity only keeps track the dependence on d and ϵ, and ignores any pre-factor that can depend on ρ which might go to infinity as $\rho \to \infty$. Indeed, it follows from $\eta \leq \bar{\eta}$ in Theorem 3.4 that $\eta \leq \frac{\log(2)}{\max_{0 \leq t \leq T} f(t)} = \frac{2\log(2)}{(b+aT)^\rho} \to 0$ as $\rho \to \infty$ so that the complexity will explode as $\rho \to \infty$.

In Table 1, we summarize the results about the iteration complexity for the examples discussed in Section 3.3. An immediate observation from Table 1 is that the iteration complexity depends on f, g and the complexity of VE-SDEs is worse than that of VP-SDEs, at least for the examples we analyze. This theoretical finding is generally consistent with Song et al. (2021), where they observed empirically that the performance of probability flow ODE samplers (with Euler or Runge-Kutta solvers) depends on the choice of forward SDEs and the sample quality for VE-SDEs is much worse than VP-SDEs for high-dimensional data. Another observation from Table 1 is that the best iteration complexity is of order $\widetilde{\mathcal{O}}(\sqrt{d}/\epsilon)$ from the examples we studied. One natural question is whether there are other choices of f, g so that the iteration complexity becomes better than $\widetilde{\mathcal{O}}(\sqrt{d}/\epsilon)$. We next show that the answer to this question is negative, if we use the result in Theorem 3.4.

Proposition 3.12. *Under the assumptions in Theorem 3.4, we further assume $\min_{t \geq 0}(g(t))^2 L(t) > 0$ and $\max_{0 \leq s \leq t} \mu(s) \leq c_1\left(\int_0^t \mu(s)\mathrm{d}s\right)^\rho + c_2$ uniformly in t for some $c_1, c_2, \rho > 0$, where $\mu(s)$ is defined in (24). We also assume that $\liminf_{T \to \infty} \int_0^T e^{-2\int_s^T f(v)\mathrm{d}v}(g(s))^2\mathrm{d}s > 0$. If we use the upper bound (12), then in order to achieve ϵ accuracy, i.e. $\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0) \leq \epsilon$, we must have $K = \widetilde{\Omega}\left(\sqrt{d}/\epsilon\right)$, where $\widetilde{\Omega}$ ignores the logarithmic dependence on ϵ and d.*

The assumptions in Proposition 3.12 are mild and one can readily check that they are satisfied for all the examples in Table 1 that achieve the iteration complexity $\widetilde{O}\left(\frac{\sqrt{d}}{\epsilon}\right)$. If we ignore the dependence on the logarithmic factors of d and ϵ, we can see from Table 1 that all the VP-SDE examples achieve the lower bound in Proposition 3.12.

4 Outline of the Proof of Theorem 3.4

We provide an outline for the proof of our main result (Theorem 3.4). At a high level, we analyze three sources of errors: (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

Table 1: Summary of the iteration complexity of the algorithm (9) for various examples in terms of ϵ and dimension d. Here f, g correspond to the drift and diffusion terms in the forward SDE (1), and a, b, 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. $\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0) \leq \epsilon$).

f	g	K	M	η
0	ae^{bt}	$\mathcal{O}\left(\frac{d^{3/2}\log(\frac{d}{\epsilon})}{\epsilon^3}\right)$	$\mathcal{O}\left(\frac{\epsilon^2}{\sqrt{d}}\right)$	$\mathcal{O}\left(\frac{\epsilon^3}{d^{3/2}}\right)$
0	$(b+at)^c$	$\mathcal{O}\left(\frac{d^{\frac{1}{(2c+1)}+\frac{3}{2}}}{\epsilon^{\frac{2}{2c+1}+3}}\right)$	$\mathcal{O}\left(\frac{\epsilon^2}{\sqrt{d}}\right)$	$\mathcal{O}\left(\frac{\epsilon^3}{d^{\frac{3}{2}}}\right)$
$\frac{b}{2}$	$\sqrt{b}$	$\mathcal{O}\left(\frac{\sqrt{d}}{\epsilon}(\log(\frac{d}{\epsilon}))^2\right)$	$\mathcal{O}\left(\frac{\epsilon}{\log(\sqrt{d}/\epsilon)}\right)$	$\mathcal{O}\left(\frac{\epsilon}{\sqrt{d}\log(\sqrt{d}/\epsilon)}\right)$
$\frac{b+at}{2}$	$\sqrt{b+at}$	$\mathcal{O}\left(\frac{\sqrt{d}}{\epsilon}(\log(\frac{d}{\epsilon}))^{\frac{3}{2}}\right)$	$\mathcal{O}\left(\frac{\epsilon}{\log(\sqrt{d}/\epsilon)}\right)$	$\mathcal{O}\left(\frac{\epsilon}{\sqrt{d}\log(\sqrt{d}/\epsilon)}\right)$
$\frac{(b+at)^\rho}{2}$	$(b+at)^{\frac{\rho}{2}}$	$\mathcal{O}\left(\frac{\sqrt{d}}{\epsilon}(\log(\frac{d}{\epsilon}))^{\frac{\rho+2}{\rho+1}}\right)$	$\mathcal{O}\left(\frac{\epsilon}{\log(\sqrt{d}/\epsilon)}\right)$	$\mathcal{O}\left(\frac{\epsilon}{\sqrt{d}\log(\sqrt{d}/\epsilon)}\right)$

continuous-time ODE (7).

First, we study the error introduced due to the initialization at $\hat{p}_T$ instead of p_T. Recall the probability flow ODE $\mathbf{y}_t$ given in (5), which has the same dynamics as $\tilde{\mathbf{x}}_t$ (defined in (2)) but with a different prior distribution $\mathbf{y}_0 \sim \hat{p}_T$ (in contrast to $\tilde{\mathbf{x}}_0 \sim p_T$). The following result bounds $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_T), p_0)$.

Proposition 4.1. *Assume p_0 is m_0-strongly-log-concave. Then, we have*

$$\mathcal{W}_2(\mathcal{L}(\mathbf{y}_T), p_0) \leq e^{-\int_0^T \mu(t)\mathrm{d}t}\|\mathbf{x}_0\|_{L_2}, \qquad (15)$$

where $\mu(t)$ is given in (24) in Appendix A.

The main challenge in analyzing the ODE $\mathbf{y}_t$ lies in studying the term $\nabla \log p_{T-t}(\mathbf{y}_t)$. In general, this term is neither linear in $\mathbf{y}_t$ nor admits a closed-form expression. However, when p_0 is strongly log-concave, it is known that that $\nabla_{\mathbf{x}} \log p_{T-t}(\mathbf{x})$ is also strongly concave (see e.g. Gao et al. (2025)). This fact allows us to establish Proposition 4.1 whose proof will be given in Appendix B.1.1.

Now we consider the algorithm (9) with iterates $(\mathbf{u}_k)$, and bound the errors due to both score estimations and discretizations. For any $k = 0, 1, 2, \ldots, K$, $\mathbf{u}_k$ has the same distribution as $\hat{\mathbf{u}}_{k\eta}$, where $\hat{\mathbf{u}}_t$ is a continuous-time process with the dynamics:

$$\frac{\mathrm{d}\hat{\mathbf{u}}_t}{\mathrm{d}t} = f(T - t)\hat{\mathbf{u}}_t$$
$$+ \frac{1}{2}(g(T - t))^2 \boldsymbol{s}_\theta \left(\hat{\mathbf{u}}_{\lfloor t/\eta\rfloor\eta}, T - \lfloor t/\eta\rfloor\eta\right), \qquad (16)$$

with the initial distribution $\hat{\mathbf{u}}_0 \sim \hat{p}_T$. We have the following result that provides an upper bound for $\|\mathbf{y}_{k\eta} - \hat{\mathbf{u}}_{k\eta}\|_{L_2}$ in terms of $\|\mathbf{y}_{(k-1)\eta} - \hat{\mathbf{u}}_{(k-1)\eta}\|_{L_2}$.

Proposition 4.2. *Assume that p_0 is m_0-strongly-log-concave, and $\nabla \log p_0$ is L_0-Lipschitz. Then, for any*

$k = 1, 2, \ldots, K$,

$$\|\mathbf{y}_{k\eta} - \hat{\mathbf{u}}_{k\eta}\|_{L_2}$$
$$\leq \gamma_{k,\eta} \cdot e^{\int_{(k-1)\eta}^{k\eta} f(T-t)\mathrm{d}t}\|\mathbf{y}_{(k-1)\eta} - \hat{\mathbf{u}}_{(k-1)\eta}\|_{L_2}$$
$$+ \frac{L_1}{2}\eta\left(1 + \|\mathbf{x}_0\|_{L_2} + \omega(T)\right)\phi_{k,\eta}$$
$$+ \frac{M}{2}\phi_{k,\eta} + \frac{\sqrt{\eta}}{2}\nu_{k,\eta}\sqrt{\psi_{k,\eta}}, \qquad (17)$$

where $\gamma_{k,\eta}$, $\phi_{k,\eta}$ and $\psi_{k,\eta}$ are defined in (23), (21) and (22) respectively, and $\omega(T)$ is defined in (32) and $\nu_{k,\eta}$ is given in (33) in Appendix A.

Proposition 4.2 provides the guarantees on how the errors due to both score estimations and discretizations propagate as the number of iterates k increases. By iterating over $k = 1, 2, \ldots, K$, we immediately get:

$$\|\mathbf{y}_{K\eta} - \hat{\mathbf{u}}_{K\eta}\|_{L_2}$$
$$\leq E_1(f, g, K, \eta, L_1) + E_2(f, g, K, \eta, M, L_1), \qquad (18)$$

where $E_1(f, g, K, \eta, L_1)$ and $E_2(f, g, K, \eta, M, L_1)$ are the discretization and score matching errors given in (13)-(14). Since $\hat{\mathbf{u}}_{k\eta}$ has the same distribution as $\mathbf{u}_k$, we have

$$\mathcal{W}_2(\mathcal{L}(\mathbf{y}_{K\eta}), \mathcal{L}(\mathbf{u}_K)) \leq \|\mathbf{y}_{K\eta} - \hat{\mathbf{u}}_{K\eta}\|_{L_2}. \qquad (19)$$

Finally, by the triangle inequality for 2-Wasserstein distance, we can decompose the 2-Wasserstein error in terms of the 2-Wasserstein error due to the initialization of the algorithm at $\hat{p}_T$ instead of p_T and the 2-Wasserstein error due to both score estimations and discretizations, we obtain:

$$\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0)$$
$$\leq \mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), \mathcal{L}(\mathbf{y}_{K\eta})) + \mathcal{W}_2(\mathcal{L}(\mathbf{y}_{K\eta}), p_0), \qquad (20)$$

where we used $T = K\eta$. Hence, Theorem 3.4 follows by applying (15), (18), (19) and (20).

Remark 4.3. Lemma 4 in Chen et al. (2023c) provides a single-step discretization analysis of the probability flow ODE in Wasserstein distance, but the (local) bound it provides is crude because it applies Gronwall's inequality, which leads to an exponential growing factor when applied recursively. In contrast, our proof leverages the strong log-concavity of the data distribution and carefully analyzes the discretization error globally. Hence, our Wasserstein analysis is novel and differs from the analysis in Chen et al. (2023c).

5   Conclusion

This paper provides the first non-asymptotic convergence analysis for a general class of probability flow ODE samplers in 2-Wasserstein distance, assuming accurate score estimates and a smooth log-concave data distribution. Our analysis provides some insights about the iteration complexity of deterministic ODE-based samplers for different choices of forward SDEs in diffusion models.

Our work serves as a first step to better understand the convergence of deterministic ODE-samplers in Wasserstein distance. It is a significant open question how to relax our current assumption of the strong-log-concave data distribution. Our proof techniques borrow the idea of synchronous coupling studied in the context of sampling from unnormalized densities using Langevin algorithms (Dalalyan and Karagulyan, 2019). To obtain Wasserstein convergence rates for using Langevin algorithm to sample from non-log-concave distributions, one may use more sophisticated coupling methods such as reflection coupling to obtain contraction rates of (Langevin) SDEs, see e.g. Eberle (2016). However, the probability flow ODE is an ODE, not an SDE, and it is not clear whether one can find an analogue of reflection coupling in the context of probability flow ODEs. Similarly, functional inequalities, another major approach to obtain convergence (typically in KL divergence) bounds for sampling with Langevin algorithms without strong-log-concavity (see e.g. Theorem 5.2.1 in Bakry et al. (2014)), are also not directly applicable to probability flow ODEs. Hence, one might need significantly different techniques to obtain Wasserstein convergence rates for ODE-based samplers without the log-concavity assumption.

In addition, while our work focuses on the sampling phase of diffusion models, another significant open problem is to investigate the training phase, i.e., understand when the score function can be accurately learned, and combine the results with the analysis of sampling to establish end-to-end guarantees for diffusion models (see e.g. Chen et al. (2023b)). We leave these investigations to the future.

References

M. S. Albergo and E. Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations*, 2022.

M. S. Albergo, N. M. Boffi, and E. Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

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

D. Bakry, I. Gentil, and M. Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer, 2014.

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*, 2024a.

J. Benton, G. Deligiannidis, and A. Doucet. Error bounds for flow matching methods. *Transactions on Machine Learning Research*, 02:1–17, 2024b.

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

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, pages 4672–4712. 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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 4462–4484. PMLR, 2023e.

X. Cheng, J. Lu, Y. Tan, and Y. Xie. Convergence of flow-based generative models via proximal gradient descent in Wasserstein space. *arXiv preprint arXiv:2310.17582*, 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, H. M. Nguyen, and L. Zhu. Wasserstein convergence guarantees for a general class of score-based generative models. *Journal of Machine Learning Research*, 2025.

A. L. Gibbs and F. E. 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.

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 *Proceedings of the 41st 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 *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=4VGEeER6W9.

Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2022.

C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. DPM-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, volume 35, pages 5775–5787, 2022.

V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov. Grad-TTS: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, volume 139, pages 8599–8608. 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.

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, S. Garg, J. Shi, and S. Ermon. Sliced score matching: A scalable approach to density and score estimation. In *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.

Y. Song, P. Dhariwal, M. Chen, and I. Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 32211–32252. PMLR, 2023.

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 Copmuting Surveys*, 56(4): 1–39, 2023.

Q. Zhang and Y. Chen. Fast sampling of diffusion models with exponential integrator. In *International Conference on Learning Representations*, 2023.

W. Zhao, L. Bai, Y. Rao, J. Zhou, and J. Lu. UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Checklist

1. For all models and algorithms presented, check if you include:

    (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]

    (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]

    (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:

    (a) Statements of the full set of assumptions of all theoretical results. [Yes]

    (b) Complete proofs of all theoretical results. [Yes]

    (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

    (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]

    (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]

    (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]

    (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

    (a) Citations of the creator If your work uses existing assets. [Not Applicable]

    (b) The license information of the assets, if applicable. [Not Applicable]

    (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

    (d) Information about consent from data providers/curators. [Not Applicable]

    (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

    (a) The full text of instructions given to participants and screenshots. [Not Applicable]

    (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to partici-
pants and the total amount spent on partici-
pant compensation. [Not Applicable]

Convergence Analysis for General Probability Flow ODEs of Diffusion Models in Wasserstein Distances

APPENDIX

The Appendix is organized as follows:

- In Appendix A, we summarize the notations given in Table 2 in the main paper, that are used in presenting the main results.

- In Appendix B, we provide the proofs of the main results of the paper.

- We present some additional technical proofs in Appendix C.

- We provide the derivation of results for various examples in Section 3.3 in the main paper and additional details in Appendix D.

A Key Quantities

In this section, we first summarize in the following Table 2 the the key quantities that play a major role in presenting the main results in the main paper.

Table 2: Summary of quantities, their interpretations and the sources

Quantities	Interpretations	Sources/References
$\bar{\eta}$ in Theorem 3.4	Upper bound for the stepsize	(27)
$\mu(t)$ in (24)	Contraction rate of $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_T), p_0)$	(35)
$L(t)$ in (26)	Lipschitz constant of $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$	Lemma B.4
$\gamma_{j,\eta}$ in (23)	Contraction rate of discretization and score-matching errors in $\mathbf{u}_j$	Proposition 4.2
$\phi_{k,\eta}$ in (21)	A component in the discretization and score-matching errors in $\mathbf{u}_j$	Theorem 3.4
$\psi_{k,\eta}$ in (22)	A component in the discretization and score-matching errors in $\mathbf{u}_j$	Theorem 3.4
$\delta_j(T-t)$ in (30)	A component in the contraction rate of discretization and score-matching errors in $\mathbf{u}_j$	Proposition B.2
$\omega(T)$ in (32)	$\sup_{0 \leq t \leq T} \|\mathbf{x}_t\|_{L_2}$	(63)
$\nu_{k,\eta}$ in (33)	Bound for $\sup_{(k-1)\eta \leq t \leq k\eta} \left\| \mathbf{y}_t - \mathbf{y}_{(k-1)\eta} \right\|_{L_2}$	Lemma B.5

Next, we provide the definitions for the key quantities in Table 2.

For any $k = 1, 2, \ldots, K$, we define:

$$\phi_{k,\eta} := \int_{(k-1)\eta}^{k\eta} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \mathrm{d}t, \tag{21}$$

$$\psi_{k,\eta} := \int_{(k-1)\eta}^{k\eta} e^{2\int_t^{k\eta} f(T-s)\mathrm{d}s} \cdot (g(T-t))^4 (L(T-t))^2 \mathrm{d}t, \tag{22}$$

and for any $j = 1, 2, \ldots, K$, we also define:

$$\gamma_{j,\eta} := 1 - \int_{(j-1)\eta}^{j\eta} \delta_j(T-t)\mathrm{d}t + \frac{L_1\eta}{2} \int_{(j-1)\eta}^{j\eta} (g(T-t))^2\mathrm{d}t, \tag{23}$$

For any $0 \leq t \leq T$, we define:

$$\mu(t) := \frac{m_0(g(t))^2}{2\left(e^{-2\int_0^t f(s)ds} + m_0 \int_0^t e^{-2\int_s^t f(v)dv}(g(s))^2 ds\right)}, \tag{24}$$

$$m(t) := \frac{(g(t))^2}{\frac{1}{m_0}e^{-2\int_0^t f(s)ds} + \int_0^t e^{-2\int_s^t f(v)dv}(g(s))^2 ds} - 2f(t), \tag{25}$$

$$L(t) := \min\left\{\left(\int_0^t e^{-2\int_s^t f(v)dv}(g(s))^2 ds\right)^{-1}, \left(e^{\int_0^t f(s)ds}\right)^2 L_0\right\}, \tag{26}$$

We also define:

$$\bar{\eta} := \min\{\bar{\eta}_1, \bar{\eta}_2\}, \tag{27}$$

$$\bar{\eta}_1 := \min\left\{\frac{\log(2)}{\max_{0 \leq t \leq T} f(t)}, \min_{0 \leq t \leq T}\left\{\frac{\frac{1}{4}(g(t))^2}{\frac{1}{m_0}e^{-2\int_0^t f(s)ds}+\int_0^t e^{-2\int_s^t f(v)dv}(g(s))^2 ds}\right\}\right\}, \tag{28}$$

$$\bar{\eta}_2 := \min_{0 \leq t \leq T}\left\{\frac{\frac{e^{-2\int_0^t f(s)ds}}{m_0} + \int_0^t e^{-2\int_s^t f(v)dv}(g(s))^2 ds}{\frac{1}{2}(g(t))^2}\right\}. \tag{29}$$

For any $k = 1, 2, \ldots, K$ and $(k-1)\eta \leq t \leq k\eta$, we define:

$$\delta_k(T-t) := \frac{\frac{1}{2}e^{-\int_{(k-1)\eta}^t f(T-s)ds}(g(T-t))^2}{\frac{1}{m_0}e^{-2\int_0^{T-t} f(s)ds} + \int_0^{T-t} e^{-2\int_s^{T-t} f(v)dv}(g(s))^2 ds} - \frac{\eta}{4}(g(T-t))^4(L(T-t))^2, \tag{30}$$

and finally, let us define:

$$\theta(T) := \sup_{0 \leq t \leq T} e^{-\frac{1}{2}\int_0^t m(T-s)ds}e^{-\int_0^T f(s)ds}\|\mathbf{x}_0\|_{L_2}, \tag{31}$$

$$\omega(T) := \sup_{0 \leq t \leq T}\left(e^{-2\int_0^t f(s)ds}\|\mathbf{x}_0\|_{L_2}^2 + d\int_0^t e^{-2\int_s^t f(v)dv}(g(s))^2 ds\right)^{1/2}, \tag{32}$$

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

$$\nu_{k,\eta} := (\theta(T) + \omega(T))\int_{(k-1)\eta}^{k\eta}\left[f(T-s) + \frac{1}{2}(g(T-s))^2 L(T-s)\right]ds$$

$$+ (L_1 T + \|\nabla \log p_0(\mathbf{0})\|)\int_{(k-1)\eta}^{k\eta} \frac{1}{2}(g(T-s))^2 ds. \tag{33}$$

B   Proofs of the Main Results

B.1   Proof of Theorem 3.4

To prove Theorem 3.4, we study the three sources of errors discussed in Section 2 for convergence analysis: (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 (7).

First, we study the error introduced due to the initialization at $\hat{p}_T$ instead of p_T. Recall the probability flow ODE $\mathbf{y}_t$ given in (5):

$$d\mathbf{y}_t = \left[f(T-t)\mathbf{y}_t + \frac{1}{2}(g(T-t))^2\nabla \log p_{T-t}(\mathbf{y}_t)\right]dt, \quad \mathbf{y}_0 \sim \hat{p}_T. \tag{34}$$

As discussed in Section 2, the distribution of $\mathbf{y}_T$ differs from p_0, because $\mathbf{y}_0 \sim \hat{p}_T \neq p_T$. The following result provides a bound on $\mathcal{W}_2(\mathcal{L}(\mathbf{y}_T), p_0)$.

Proposition B.1 (Restatement of Proposition 4.1). *Assume that p_0 is m_0-strongly-log-concave. Then, we have*

$$\mathcal{W}_2(\mathcal{L}(\mathbf{y}_T), p_0) \le e^{-\int_0^T \mu(t)\mathrm{d}t}\|\mathbf{x}_0\|_{L_2}, \tag{35}$$

where $\mu(t)$ is given in (24).

Notice that the term $\|\mathbf{x}_0\|_{L_2}$ in Proposition B.1 is finite since Assumption 3.1 implies that $\mathbf{x}_0 \sim p_0$ is L_2-integrable (see e.g. Lemma 11 in Gürbüzbalaban et al. (2021)).

The key idea of the proof of Proposition B.1 is to observe that when p_0 is strongly log-concave, the term $\nabla_{\mathbf{x}} \log p_{T-t}(\mathbf{x})$ is also strongly concave (see e.g. Gao et al. (2025)). This fact allows us to establish Proposition B.1. The proof of Proposition B.1 will be given in Section B.1.1.

Now we consider the algorithm (9) with iterates $(\mathbf{u}_k)$, and bound the errors due to score estimations and discretizations together. For any $k = 0, 1, 2, \ldots, K$, $\mathbf{u}_k$ has the same distribution as $\hat{\mathbf{u}}_{k\eta}$, where $\hat{\mathbf{u}}_t$ is a continuous-time process with the dynamics:

$$d\hat{\mathbf{u}}_t = \left[f(T-t)\hat{\mathbf{u}}_t + \frac{1}{2}(g(T-t))^2 \boldsymbol{s}_\theta \left(\hat{\mathbf{u}}_{\lfloor t/\eta \rfloor \eta}, T - \lfloor t/\eta \rfloor \eta \right) \right] \mathrm{d}t, \tag{36}$$

with the initial distribution $\hat{\mathbf{u}}_0 \sim \hat{p}_T$. We have the following result that provides an upper bound for $\|\mathbf{y}_{k\eta} - \hat{\mathbf{u}}_{k\eta}\|_{L_2}$ in terms of $\|\mathbf{y}_{(k-1)\eta} - \hat{\mathbf{u}}_{(k-1)\eta}\|_{L_2}$. This result plays a key role in the proof of Theorem 3.4.

Proposition B.2 (Restatement of Proposition 4.2). *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 $k = 1, 2, \ldots, K$,*

$$\|\mathbf{y}_{k\eta} - \hat{\mathbf{u}}_{k\eta}\|_{L_2} \le \left(1 - \int_{(k-1)\eta}^{k\eta} \delta_k(T-t)\mathrm{d}t + \frac{L_1}{2}\eta \int_{(k-1)\eta}^{k\eta} (g(T-t))^2 \mathrm{d}t \right)$$

$$\cdot e^{\int_{(k-1)\eta}^{k\eta} f(T-t)\mathrm{d}t}\|\mathbf{y}_{(k-1)\eta} - \hat{\mathbf{u}}_{(k-1)\eta}\|_{L_2}$$

$$+ \frac{L_1}{2}\eta \left(1 + \|\mathbf{x}_0\|_{L_2} + \omega(T) \right) \int_{(k-1)\eta}^{k\eta} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s}(g(T-t))^2 \mathrm{d}t$$

$$+ \frac{M}{2} \int_{(k-1)\eta}^{k\eta} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s}(g(T-t))^2 \mathrm{d}t$$

$$+ \sqrt{\eta}\nu_{k,\eta} \left(\int_{(k-1)\eta}^{k\eta} \left[\frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s}(g(T-t))^2 L(T-t) \right]^2 \mathrm{d}t \right)^{1/2}, \tag{37}$$

where $\delta_k(t)$, $0 \le t \le T$, is defined in (30), $\omega(T)$ is defined in (32) and $\nu_{k,\eta}$ is given in (33).

We remark that the coefficient in front of the term $\left\|\mathbf{y}_{(k-1)\eta} - \hat{\mathbf{u}}_{(k-1)\eta}\right\|_{L_2}$ in (37) lies in between zero and one. Indeed, we can see that assumption $\eta \le \bar{\eta} := \min(\bar{\eta}_1, \bar{\eta}_2)$ in Theorem 3.4 implies that $\eta \le \bar{\eta}_1$, where $\bar{\eta}_1$ is defined in (28) which yields that

$$\eta \le \min_{0 \le t \le T} \left\{ \frac{\frac{\frac{1}{2}e^{-\eta \max_{0 \le t \le T} f(t)}(g(t))^2}{\frac{1}{m_0}e^{-2\int_0^t f(s)\mathrm{d}s} + \int_0^t e^{-2\int_s^t f(v)\mathrm{d}v}(g(s))^2 \mathrm{d}s}}{\frac{1}{4}(g(t))^4(L(t))^2 + \frac{L_1}{2}(g(t))^2} \right\}, \tag{38}$$

and assumption $\eta \le \bar{\eta} := \min(\bar{\eta}_1, \bar{\eta}_2)$ in Theorem 3.4 implies that $\eta \le \bar{\eta}_2$, where $\bar{\eta}_2$ is defined in (29) which yields that

$$\eta \le \min_{0 \le t \le T} \left\{ \frac{\frac{1}{m_0}e^{-2\int_0^t f(s)\mathrm{d}s} + \int_0^t e^{-2\int_s^t f(v)\mathrm{d}v}(g(s))^2 \mathrm{d}s}{\frac{1}{2}e^{-\eta \min_{0 \le t \le T} f(t)}(g(t))^2} \right\}, \tag{39}$$

and it follows from (38)-(39) and the definition of $\delta_j(t)$ in (30) that $\delta_j(T-t) \ge \frac{L_1}{2}\eta(g(T-t))^2$ for every $j = 1, 2, \ldots, K$ and $(j-1)\eta \le t \le j\eta$ and $\eta \max_{(j-1)\eta \le t \le j\eta} \delta_j(t) < 1$ for every $j = 1, 2, \ldots, K$ such that for any $j = 1, 2, \ldots, K$,

$$0 \le 1 - \int_{(j-1)\eta}^{j\eta} \delta_j(T-t)\mathrm{d}t + \frac{L_1}{2}\eta \int_{(j-1)\eta}^{j\eta} (g(T-t))^2 \mathrm{d}t \le 1.$$

Now we are ready to prove Theorem 3.4.

Proof of Theorem 3.4. Since $\hat{\mathbf{u}}_{k\eta}$ has the same distribution as $\mathbf{u}_k$, by applying (37) recursively, we have

$$\mathcal{W}_2(\mathcal{L}(\mathbf{y}_{K\eta}), \mathcal{L}(\mathbf{u}_K)) \leq \|\mathbf{y}_{K\eta} - \hat{\mathbf{u}}_{K\eta}\|_{L_2}$$

$$\leq \sum_{k=1}^{K} \prod_{j=k+1}^{K} \left(1 - \int_{(j-1)\eta}^{j\eta} \delta_j(T-t)\mathrm{d}t + \frac{L_1}{2}\eta \int_{(j-1)\eta}^{j\eta} (g(T-t))^2\mathrm{d}t\right)$$

$$\cdot e^{\int_{k\eta}^{K\eta} f(T-t)\mathrm{d}t} \left(\frac{L_1}{2}\eta\left(1 + \|\mathbf{x}_0\|_{L_2} + \omega(T)\right)\int_{(k-1)\eta}^{k\eta} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s}(g(T-t))^2\mathrm{d}t\right.$$

$$+ \frac{M}{2}\int_{(k-1)\eta}^{k\eta} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s}(g(T-t))^2\mathrm{d}t$$

$$\left.+ \sqrt{\eta}\nu_{k,\eta}\left(\int_{(k-1)\eta}^{k\eta}\left[\frac{1}{2}e^{\int_t^{k\eta} f(T-s)\mathrm{d}s}(g(T-t))^2 L(T-t)\right]^2 \mathrm{d}t\right)^{1/2}\right).$$

Moreover, we recall that $T = K\eta$ and by triangle inequality for 2-Wasserstein distance,

$$\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0) \leq \mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), \mathcal{L}(\mathbf{y}_{K\eta})) + \mathcal{W}_2(\mathcal{L}(\mathbf{y}_{K\eta}), p_0). \tag{40}$$

By applying Proposition B.1 and (40), we get

$$\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0) \leq e^{-\int_0^{K\eta} \mu(t)\mathrm{d}t} \cdot \|\mathbf{x}_0\|_{L_2} + E_1(f, g, K, \eta, L_1) + E_2(f, g, K, \eta, M, L_1), \tag{41}$$

where $\mu(t)$ is given in (24) and we recall from (13)-(14) that

$$E_1(f, g, K, \eta, L_1) := \sum_{k=1}^{K} \prod_{j=k+1}^{K} \gamma_{j,\eta} \cdot e^{\int_{k\eta}^{K\eta} f(T-t)\mathrm{d}t} \cdot \left(\frac{L_1}{2}\eta\left(1 + \|\mathbf{x}_0\|_{L_2} + \omega(T)\right)\phi_{k,\eta} + \frac{\sqrt{\eta}}{2}\nu_{k,\eta}\sqrt{\psi_{k,\eta}}\right), \tag{42}$$

$$E_2(f, g, K, \eta, M, L_1) := \sum_{k=1}^{K} \prod_{j=k+1}^{K} \gamma_{j,\eta} \cdot e^{\int_{k\eta}^{K\eta} f(T-t)\mathrm{d}t} \cdot \frac{M}{2}\phi_{k,\eta}, \tag{43}$$

where $\phi_{k,\eta}$ is given in (21), $\psi_{k,\eta}$ is given in (22), $\gamma_{j,\eta}$ is given in (23), $L(t)$ is given in (26), $\delta_j(T-t)$ is defined in (30), $\omega(T)$ is defined in (32) and $\nu_{k,\eta}$ is given in (33). The proof is complete. $\qquad\square$

B.1.1 Proof of Proposition B.1

Before we proceed to the proof of Proposition B.1, let us first introduce a technical lemma.

Lemma B.3. *It holds that:*

$$\mathcal{W}_2(p_T, \hat{p}_T) \leq e^{-\int_0^T f(s)\mathrm{d}s}\|\mathbf{x}_0\|_{L_2}. \tag{44}$$

Proof of Lemma B.3. We recall that p_T is the distribution of $\mathbf{x}_T$ which has the expression (see (4))

$$\mathbf{x}_T = e^{-\int_0^T f(s)\mathrm{d}s}\mathbf{x}_0 + \int_0^T e^{-\int_s^T f(v)\mathrm{d}v}g(s)\mathrm{d}\mathbf{B}_s, \tag{45}$$

and $\hat{p}_T$ (see (3)) is the distribution of

$$\hat{\mathbf{x}}_T = \int_0^T e^{-\int_s^T f(v)\mathrm{d}v}g(s)\mathrm{d}\mathbf{B}_s. \tag{46}$$

Therefore, it follows from (45) and (46) that

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

This completes the proof. $\qquad\square$

Now, we are ready to prove Proposition B.1.

Proof of Proposition B.1. We recall that

$$d\tilde{\mathbf{x}}_t = \left[f(T-t)\tilde{\mathbf{x}}_t + \frac{1}{2}(g(T-t))^2 \nabla \log p_{T-t}(\tilde{\mathbf{x}}_t) \right] dt, \tag{47}$$

with the initial distribution $\tilde{\mathbf{x}}_0 \sim p_T$ and

$$d\mathbf{y}_t = \left[f(T-t)\mathbf{y}_t + \frac{1}{2}(g(T-t))^2 \nabla \log p_{T-t}(\mathbf{y}_t) \right] dt,$$

with the initial distribution $\mathbf{y}_0 \sim \hat{p}_T$.

It is proved in Gao et al. (2025) that $\log p_{T-t}(\mathbf{x})$ is $a(T-t)$-strongly-concave, where

$$a(T-t) := \frac{1}{\frac{1}{m_0}e^{-2\int_0^{T-t} f(s)ds} + \int_0^{T-t} e^{-2\int_s^{T-t} f(v)dv}(g(s))^2 ds}. \tag{48}$$

Next, let us recall from (25) the definition of $m(T-t)$:

$$m(T-t) := (g(T-t))^2 a(T-t) - 2f(T-t), \qquad 0 \le t \le T, \tag{49}$$

where $a(T-t)$ is defined in (48). We can compute that

$$d\left(\|\tilde{\mathbf{x}}_t - \mathbf{y}_t\|^2 e^{\int_0^t m(T-s)ds} \right)$$
$$= m(T-t)e^{\int_0^t m(T-s)ds}\|\tilde{\mathbf{x}}_t - \mathbf{y}_t\|^2 dt + 2e^{\int_0^t m(T-s)ds}\langle \tilde{\mathbf{x}}_t - \mathbf{y}_t, d\tilde{\mathbf{x}}_t - d\mathbf{y}_t \rangle$$
$$= m(T-t)e^{\int_0^t m(T-s)ds}\|\tilde{\mathbf{x}}_t - \mathbf{y}_t\|^2 dt + 2e^{\int_0^t m(T-s)ds}\langle \tilde{\mathbf{x}}_t - \mathbf{y}_t, f(T-t)(\tilde{\mathbf{x}}_t - \mathbf{y}_t) \rangle dt$$
$$\quad + 2e^{\int_0^t m(T-s)ds}\left\langle \tilde{\mathbf{x}}_t - \mathbf{y}_t, \frac{1}{2}(g(T-t))^2 \left(\nabla \log p_{T-t}(\tilde{\mathbf{x}}_t) - \nabla \log p_{T-t}(\mathbf{y}_t) \right) \right\rangle dt$$
$$\le e^{\int_0^t m(T-s)ds} \left(m(T-t) + 2f(T-t) - (g(T-t))^2 a(T-t) \right) \|\tilde{\mathbf{x}}_t - \mathbf{y}_t\|^2 dt$$
$$= 0.$$

This implies that

$$\|\tilde{\mathbf{x}}_t - \mathbf{y}_t\|^2 e^{\int_0^t m(T-s)ds} \le \|\tilde{\mathbf{x}}_0 - \mathbf{y}_0\|^2, \tag{50}$$

so that

$$\mathbb{E}\|\tilde{\mathbf{x}}_T - \mathbf{y}_T\|^2 \le e^{-\int_0^T m(T-s)ds}\mathbb{E}\|\tilde{\mathbf{x}}_0 - \mathbf{y}_0\|^2. \tag{51}$$

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

Next, we recall from Lemma B.3 that

$$\mathcal{W}_2(p_T, \hat{p}_T) \le e^{-\int_0^T f(s)ds}\|\mathbf{x}_0\|_{L_2}. \tag{52}$$

By combining (51) with (52), we conclude that

$$\mathcal{W}_2^2(\mathcal{L}(\mathbf{y}_T), p_0) = \mathcal{W}_2(\mathcal{L}(\mathbf{y}_T), \mathcal{L}(\tilde{\mathbf{x}}_T)) \le \mathbb{E}\|\tilde{\mathbf{x}}_T - \mathbf{y}_T\|^2$$
$$\le e^{-\int_0^T m(T-s)ds}\mathcal{W}_2^2(p_T, \hat{p}_T)$$
$$\le e^{-\int_0^T m(s)ds}e^{-2\int_0^T f(s)ds}\|\mathbf{x}_0\|_{L_2}^2 = e^{-2\int_0^T \mu(t)dt}\|\mathbf{x}_0\|_{L_2}^2,$$

where

$$\mu(t) = f(t) + \frac{m(t)}{2} = \frac{m_0(g(t))^2}{2\left(e^{-2\int_0^t f(s)ds} + m_0 \int_0^t e^{-2\int_s^t f(v)dv}(g(s))^2 ds \right)},$$

and we have used (49). The proof is complete. $\qquad\qquad\square$

B.1.2 Proof of Proposition B.2

We first state a key technical lemma, which will be used in the proof of Proposition B.2.

Lemma B.4 (Gao et al. (2025)). *Suppose that Assumption 3.1 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 (26).*

Proof of Proposition B.2. First, we recall that for any $(k-1)\eta \leq t \leq k\eta$,

$$\mathbf{y}_t = \mathbf{y}_{(k-1)\eta} + \int_{(k-1)\eta}^{t} \left[f(T-s)\mathbf{y}_s + \frac{1}{2}(g(T-s))^2 \nabla \log p_{T-s}(\mathbf{y}_s) \right] \mathrm{d}s,$$

$$\hat{\mathbf{u}}_t = \hat{\mathbf{u}}_{(k-1)\eta} + \int_{(k-1)\eta}^{t} \left[f(T-s)\hat{\mathbf{u}}_s + \frac{1}{2}(g(T-s))^2 \boldsymbol{s}_\theta \left(\hat{\mathbf{u}}_{(k-1)\eta}, T - (k-1)\eta \right) \right] \mathrm{d}s,$$

which implies that

$$\mathbf{y}_{k\eta} = e^{\int_{(k-1)\eta}^{k\eta} f(T-t)\mathrm{d}t} \mathbf{y}_{(k-1)\eta} + \frac{1}{2} \int_{(k-1)\eta}^{k\eta} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \nabla \log p_{T-t}(\mathbf{y}_t) \mathrm{d}t,$$

$$\hat{\mathbf{u}}_{k\eta} = e^{\int_{(k-1)\eta}^{k\eta} f(T-t)\mathrm{d}t} \hat{\mathbf{u}}_{(k-1)\eta} + \frac{1}{2} \int_{(k-1)\eta}^{k\eta} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \boldsymbol{s}_\theta \left(\hat{\mathbf{u}}_{(k-1)\eta}, T - (k-1)\eta \right) \mathrm{d}t.$$

It follows that

$$\begin{aligned}
&\mathbf{y}_{k\eta} - \hat{\mathbf{u}}_{k\eta} \\
&= e^{\int_{(k-1)\eta}^{k\eta} f(T-t)\mathrm{d}t} \left(\mathbf{y}_{(k-1)\eta} - \hat{\mathbf{u}}_{(k-1)\eta} \right) \\
&\quad + \int_{(k-1)\eta}^{k\eta} \frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \left(\nabla \log p_{T-t}(\mathbf{y}_{(k-1)\eta}) - \nabla \log p_{T-t}\left(\hat{\mathbf{u}}_{(k-1)\eta} \right) \right) \mathrm{d}t \\
&\quad + \int_{(k-1)\eta}^{k\eta} \frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \left(\nabla \log p_{T-t}(\mathbf{y}_t) - \nabla \log p_{T-t}(\mathbf{y}_{(k-1)\eta}) \right) \mathrm{d}t \\
&\quad + \int_{(k-1)\eta}^{k\eta} \frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \left(\nabla \log p_{T-t}\left(\hat{\mathbf{u}}_{(k-1)\eta} \right) - \boldsymbol{s}_\theta \left(\hat{\mathbf{u}}_{(k-1)\eta}, T - (k-1)\eta \right) \right) \mathrm{d}t.
\end{aligned}$$

This implies that

$$\begin{aligned}
&\left\| \mathbf{y}_{k\eta} - \hat{\mathbf{u}}_{k\eta} \right\|_{L_2} \\
&\leq \left\| e^{\int_{(k-1)\eta}^{k\eta} f(T-t)\mathrm{d}t} \left(\mathbf{y}_{(k-1)\eta} - \hat{\mathbf{u}}_{(k-1)\eta} \right) \right. \\
&\qquad \left. + \int_{(k-1)\eta}^{k\eta} \frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \left(\nabla \log p_{T-t}(\mathbf{y}_{(k-1)\eta}) - \nabla \log p_{T-t}\left(\hat{\mathbf{u}}_{(k-1)\eta} \right) \right) \mathrm{d}t \right\|_{L_2} \\
&\quad + \left\| \int_{(k-1)\eta}^{k\eta} \frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \left(\nabla \log p_{T-t}(\mathbf{y}_t) - \nabla \log p_{T-t}(\mathbf{y}_{(k-1)\eta}) \right) \mathrm{d}t \right\|_{L_2} \\
&\quad + \left\| \int_{(k-1)\eta}^{k\eta} \frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \right. \\
&\qquad\qquad \left. \cdot \left(\nabla \log p_{T-t}\left(\hat{\mathbf{u}}_{(k-1)\eta} \right) - \boldsymbol{s}_\theta \left(\hat{\mathbf{u}}_{(k-1)\eta}, T - (k-1)\eta \right) \right) \mathrm{d}t \right\|_{L_2}.
\end{aligned} \tag{53}$$

Next, we provide upper bounds for the three terms in (53).

Bounding the first term in (53)**.** We can compute that

$$
\left\| e^{\int_{(k-1)\eta}^{k\eta} f(T-t)\mathrm{d}t} \left(\mathbf{y}_{(k-1)\eta} - \hat{\mathbf{u}}_{(k-1)\eta} \right) \right.
$$

$$
\left. + \int_{(k-1)\eta}^{k\eta} \frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \left(\nabla \log p_{T-t}(\mathbf{y}_{(k-1)\eta}) - \nabla \log p_{T-t}\left(\hat{\mathbf{u}}_{(k-1)\eta} \right) \right) \mathrm{d}t \right\|^2
$$

$$
= e^{2 \int_{(k-1)\eta}^{k\eta} f(T-t)\mathrm{d}t} \left\| \mathbf{y}_{(k-1)\eta} - \hat{\mathbf{u}}_{(k-1)\eta} \right\|^2
$$

$$
+ \left\| \int_{(k-1)\eta}^{k\eta} \frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \left(\nabla \log p_{T-t}(\mathbf{y}_{(k-1)\eta}) - \nabla \log p_{T-t}\left(\hat{\mathbf{u}}_{(k-1)\eta} \right) \right) \mathrm{d}t \right\|^2
$$

$$
+ 2 \int_{(k-1)\eta}^{k\eta} \left\langle e^{\int_{(k-1)\eta}^{k\eta} f(T-t)\mathrm{d}t} \left(\mathbf{y}_{(k-1)\eta} - \hat{\mathbf{u}}_{(k-1)\eta} \right), \right.
$$

$$
\left. \frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \left(\nabla \log p_{T-t}(\mathbf{y}_{(k-1)\eta}) - \nabla \log p_{T-t}\left(\hat{\mathbf{u}}_{(k-1)\eta} \right) \right) \right\rangle \mathrm{d}t.
$$

We know from Gao et al. (2025) that $\log p_{T-t}(\mathbf{x})$ is $a(T-t)$-strongly-concave, where $a(T-t)$ is given in (48). Hence we have

$$
\left\| e^{\int_{(k-1)\eta}^{k\eta} f(T-t)\mathrm{d}t} \left(\mathbf{y}_{(k-1)\eta} - \hat{\mathbf{u}}_{(k-1)\eta} \right) \right.
$$

$$
\left. + \int_{(k-1)\eta}^{k\eta} \frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \left(\nabla \log p_{T-t}(\mathbf{y}_{(k-1)\eta}) - \nabla \log p_{T-t}\left(\hat{\mathbf{u}}_{(k-1)\eta} \right) \right) \mathrm{d}t \right\|^2
$$

$$
\leq \left(1 - \int_{(k-1)\eta}^{k\eta} m_k(T-t)\mathrm{d}t \right) e^{2 \int_{(k-1)\eta}^{k\eta} f(T-t)\mathrm{d}t} \left\| \mathbf{y}_{(k-1)\eta} - \hat{\mathbf{u}}_{(k-1)\eta} \right\|^2
$$

$$
+ \left(\int_{(k-1)\eta}^{k\eta} \frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 L(T-t) \left\| \mathbf{y}_{(k-1)\eta} - \hat{\mathbf{u}}_{(k-1)\eta} \right\| \mathrm{d}t \right)^2
$$

$$
\leq \left(1 - \int_{(k-1)\eta}^{k\eta} m_k(T-t)\mathrm{d}t + \frac{\eta}{2} \int_{(k-1)\eta}^{k\eta} (g(T-t))^4 (L(T-t))^2 \mathrm{d}t \right)
$$

$$
\cdot e^{2 \int_{(k-1)\eta}^{k\eta} f(T-t)\mathrm{d}t} \left\| \mathbf{y}_{(k-1)\eta} - \hat{\mathbf{u}}_{(k-1)\eta} \right\|^2,
$$

where we applied Cauchy-Schwartz inequality and Lemma B.4, and $m_k(T-t)$ is defined as:

$$
m_k(T-t) := e^{-\int_{(k-1)\eta}^t f(T-s)\mathrm{d}s} (g(T-t))^2 a(T-t), \qquad (k-1)\eta \leq t \leq k\eta, \tag{54}
$$

for every $k = 1, 2, \ldots, K$. Hence, we conclude that

$$
\left\| e^{\int_{(k-1)\eta}^{k\eta} f(T-t)\mathrm{d}t} \left(\mathbf{y}_{(k-1)\eta} - \hat{\mathbf{u}}_{(k-1)\eta} \right) \right.
$$

$$
\left. + \int_{(k-1)\eta}^{k\eta} \frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \left(\nabla \log p_{T-t}\left(\mathbf{y}_{(k-1)\eta} \right) - \nabla \log p_{T-t}\left(\hat{\mathbf{u}}_{(k-1)\eta} \right) \right) \mathrm{d}t \right\|_{L_2}
$$

$$
\leq \left(1 - \int_{(k-1)\eta}^{k\eta} \delta_k(T-t)\mathrm{d}t \right) e^{\int_{(k-1)\eta}^{k\eta} f(T-t)\mathrm{d}t} \left\| \mathbf{y}_{(k-1)\eta} - \hat{\mathbf{u}}_{(k-1)\eta} \right\|_{L_2}, \tag{55}
$$

where we used the inequality $\sqrt{1-x} \leq 1 - \frac{x}{2}$ for any $0 \leq x \leq 1$ and the definition of $\delta_k(T-t)$ in (30) which can be rewritten as

$$
\delta_k(T-t) := \frac{1}{2} e^{-\int_{(k-1)\eta}^t f(T-s)\mathrm{d}s} (g(T-t))^2 a(T-t) - \frac{\eta}{4} (g(T-t))^4 (L(T-t))^2, \quad (k-1)\eta \leq t \leq k\eta,
$$

where $a(T - t)$ is given in (48).

Bounding the second term in (53). Using Lemma B.4, we can compute that

$$\left\| \int_{(k-1)\eta}^{k\eta} \frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \left(\nabla \log p_{T-t}(\mathbf{y}_t) - \nabla \log p_{T-t}(\mathbf{y}_{(k-1)\eta}) \right) \mathrm{d}t \right\|^2$$

$$\leq \left(\int_{(k-1)\eta}^{k\eta} \frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 L(T-t) \|\mathbf{y}_t - \mathbf{y}_{(k-1)\eta}\| \mathrm{d}t \right)^2$$

$$\leq \eta \int_{(k-1)\eta}^{k\eta} \left[\frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 L(T-t) \right]^2 \|\mathbf{y}_t - \mathbf{y}_{(k-1)\eta}\|^2 \mathrm{d}t,$$

which implies that

$$\left\| \int_{(k-1)\eta}^{k\eta} \frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \left(\nabla \log p_{T-t}(\mathbf{y}_t) - \nabla \log p_{T-t}\left(\mathbf{y}_{(k-1)\eta}\right) \right) \mathrm{d}t \right\|_{L_2}$$

$$\leq \left(\mathbb{E} \left[\eta \int_{(k-1)\eta}^{k\eta} \left[\frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 L(T-t) \right]^2 \|\mathbf{y}_t - \mathbf{y}_{(k-1)\eta}\|^2 \mathrm{d}t \right] \right)^{1/2}$$

$$\leq \left(\eta \int_{(k-1)\eta}^{k\eta} \left[\frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 L(T-t) \right]^2 \mathrm{d}t \cdot \sup_{(k-1)\eta \leq t \leq k\eta} \mathbb{E} \|\mathbf{y}_t - \mathbf{y}_{(k-1)\eta}\|^2 \right)^{1/2}$$

$$= \sqrt{\eta} \left(\int_{(k-1)\eta}^{k\eta} \left[\frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 L(T-t) \right]^2 \mathrm{d}t \right)^{1/2} \sup_{(k-1)\eta \leq t \leq k\eta} \left\| \mathbf{y}_t - \mathbf{y}_{(k-1)\eta} \right\|_{L_2}. \qquad (56)$$

Bounding the third term in (53). We notice that

$$\left\| \int_{(k-1)\eta}^{k\eta} \frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \left(\nabla \log p_{T-t}(\hat{\mathbf{u}}_{(k-1)\eta}) - \boldsymbol{s}_\theta \left(\hat{\mathbf{u}}_{(k-1)\eta}, T - (k-1)\eta \right) \right) \mathrm{d}t \right\|_{L_2}$$

$$\leq \left\| \int_{(k-1)\eta}^{k\eta} \frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \left(\nabla \log p_{T-(k-1)\eta}(\hat{\mathbf{u}}_{(k-1)\eta}) - \boldsymbol{s}_\theta \left(\hat{\mathbf{u}}_{(k-1)\eta}, T - (k-1)\eta \right) \right) \mathrm{d}t \right\|_{L_2}$$

$$+ \left\| \int_{(k-1)\eta}^{k\eta} \frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \left(\nabla \log p_{T-t}\left(\hat{\mathbf{u}}_{(k-1)\eta}\right) - \nabla \log p_{T-(k-1)\eta}\left(\hat{\mathbf{u}}_{(k-1)\eta}\right) \right) \mathrm{d}t \right\|_{L_2}.$$

By Assumption 3.3, we have

$$\left\| \int_{(k-1)\eta}^{k\eta} \frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \left(\nabla \log p_{T-(k-1)\eta}(\hat{\mathbf{u}}_{(k-1)\eta}) - \boldsymbol{s}_\theta \left(\hat{\mathbf{u}}_{(k-1)\eta}, T - (k-1)\eta \right) \right) \mathrm{d}t \right\|_{L_2}$$

$$\leq \frac{M}{2} \int_{(k-1)\eta}^{k\eta} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \mathrm{d}t. \qquad (57)$$

Moreover, by Assumption 3.2, we have

$$\left\| \int_{(k-1)\eta}^{k\eta} \frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \left(\nabla \log p_{T-t}\left(\hat{\mathbf{u}}_{(k-1)\eta}\right) - \nabla \log p_{T-(k-1)\eta}\left(\hat{\mathbf{u}}_{(k-1)\eta}\right) \right) \mathrm{d}t \right\|_{L_2}$$

$$\leq \int_{(k-1)\eta}^{k\eta} \frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \left\| \nabla \log p_{T-t}\left(\hat{\mathbf{u}}_{(k-1)\eta}\right) - \nabla \log p_{T-(k-1)\eta}\left(\hat{\mathbf{u}}_{(k-1)\eta}\right) \right\|_{L_2} \mathrm{d}t$$

$$\leq \int_{(k-1)\eta}^{k\eta} \frac{1}{2} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 L_1 \eta \left(1 + \left\| \hat{\mathbf{u}}_{(k-1)\eta} \right\|_{L_2} \right) \mathrm{d}t$$

$$\leq \frac{L_1}{2} \eta \left(1 + \left\| \mathbf{y}_{(k-1)\eta} - \hat{\mathbf{u}}_{(k-1)\eta} \right\|_{L_2} + \left\| \mathbf{y}_{(k-1)\eta} \right\|_{L_2} \right) \int_{(k-1)\eta}^{k\eta} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s} (g(T-t))^2 \mathrm{d}t. \qquad (58)$$

Furthermore, we can compute that

$$\left\|\mathbf{y}_{(k-1)\eta}\right\|_{L_2} \le \left\|\mathbf{y}_{(k-1)\eta} - \tilde{\mathbf{x}}_{(k-1)\eta}\right\|_{L_2} + \left\|\tilde{\mathbf{x}}_{(k-1)\eta}\right\|_{L_2}, \tag{59}$$

where $\tilde{\mathbf{x}}_t$ is defined in (2). Moreover, by (50) in the proof of Proposition B.1, we have

$$\left\|\mathbf{y}_{(k-1)\eta} - \tilde{\mathbf{x}}_{(k-1)\eta}\right\|_{L_2} \le \left(\mathbb{E}\|\tilde{\mathbf{x}}_0 - \mathbf{y}_0\|^2\right)^{1/2} = e^{-\int_0^T f(s)\mathrm{d}s}\|\mathbf{x}_0\|_{L_2} \le \|\mathbf{x}_0\|_{L_2}, \tag{60}$$

where we applied (4) to obtain the equality in the above equation. Moreover, since $\tilde{\mathbf{x}}_t = \mathbf{x}_{T-t}$ in distribution for any $t \in [0, T]$, we have

$$\left\|\tilde{\mathbf{x}}_{(k-1)\eta}\right\|_{L_2} = \left\|\mathbf{x}_{T-(k-1)\eta}\right\|_{L_2} \le \sup_{0 \le t \le T} \|\mathbf{x}_t\|_{L_2} =: \omega(T). \tag{61}$$

Next, let us show that $\omega(T)$ can be computed as given by the formula in (32). By equation (4), we have

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

By taking expectations, we obtain

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

which implies that

$$\mathbb{E}\|\mathbf{x}_t\|^2 = e^{-2\int_0^t f(s)\mathrm{d}s}\mathbb{E}\|\mathbf{x}_0\|^2 + d\int_0^t e^{-2\int_s^t f(v)\mathrm{d}v}(g(s))^2\mathrm{d}s. \tag{62}$$

Therefore, we conclude that

$$\omega(T) = \sup_{0 \le t \le T} \|\mathbf{x}_t\|_{L_2} = \sup_{0 \le t \le T} \left(e^{-2\int_0^t f(s)\mathrm{d}s}\|\mathbf{x}_0\|_{L_2}^2 + d\int_0^t e^{-2\int_s^t f(v)\mathrm{d}v}(g(s))^2\mathrm{d}s\right)^{1/2}. \tag{63}$$

Therefore, by applying (58), (59), (60) and (61), we have

$$\left\|\int_{(k-1)\eta}^{k\eta} \frac{1}{2}e^{\int_t^{k\eta} f(T-s)\mathrm{d}s}(g(T-t))^2\left(\nabla\log p_{T-t}\left(\hat{\mathbf{u}}_{(k-1)\eta}\right) - \nabla\log p_{T-(k-1)\eta}\left(\hat{\mathbf{u}}_{(k-1)\eta}\right)\right)\mathrm{d}t\right\|_{L_2}$$
$$\le \frac{L_1}{2}\eta\left\|\mathbf{y}_{(k-1)\eta} - \hat{\mathbf{u}}_{(k-1)\eta}\right\|_{L_2}\int_{(k-1)\eta}^{k\eta} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s}(g(T-t))^2\mathrm{d}t$$
$$+ \frac{L_1}{2}\eta\left(1 + \|\mathbf{x}_0\|_{L_2} + \omega(T)\right)\int_{(k-1)\eta}^{k\eta} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s}(g(T-t))^2\mathrm{d}t.$$

It follows that the third term in (53) is upper bounded by

$$\frac{L_1}{2}\eta\left\|\mathbf{y}_{(k-1)\eta} - \hat{\mathbf{u}}_{(k-1)\eta}\right\|_{L_2}\int_{(k-1)\eta}^{k\eta} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s}(g(T-t))^2\mathrm{d}t$$
$$+ \frac{L_1}{2}\eta\left(1 + \|\mathbf{x}_0\|_{L_2} + \omega(T)\right)\int_{(k-1)\eta}^{k\eta} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s}(g(T-t))^2\mathrm{d}t$$
$$+ \frac{M}{2}\int_{(k-1)\eta}^{k\eta} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s}(g(T-t))^2\mathrm{d}t. \tag{64}$$

Bounding (53). On combining (55), (56) and (64), we conclude that

$$
\begin{aligned}
&\|\mathbf{y}_{k\eta} - \hat{\mathbf{u}}_{k\eta}\|_{L_2}^2 \\
&\leq \Bigg\{ \left(1 - \int_{(k-1)\eta}^{k\eta} \delta_k(T-t)\mathrm{d}t + \frac{L_1}{2}\eta \int_{(k-1)\eta}^{k\eta} (g(T-t))^2\mathrm{d}t \right) e^{\int_{(k-1)\eta}^{k\eta} f(T-t)\mathrm{d}t} \left\|\mathbf{y}_{(k-1)\eta} - \hat{\mathbf{u}}_{(k-1)\eta}\right\|_{L_2} \\
&\quad + \frac{L_1}{2}\eta\left(1 + \|\mathbf{x}_0\|_{L_2} + \omega(T)\right) \int_{(k-1)\eta}^{k\eta} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s}(g(T-t))^2\mathrm{d}t \\
&\quad + \frac{M}{2} \int_{(k-1)\eta}^{k\eta} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s}(g(T-t))^2\mathrm{d}t \\
&\quad + \sqrt{\eta}\left(\int_{(k-1)\eta}^{k\eta} \left[\frac{1}{2}e^{\int_t^{k\eta} f(T-s)\mathrm{d}s}(g(T-t))^2 L(T-t)\right]^2 \mathrm{d}t \right)^{1/2} \sup_{(k-1)\eta \leq t \leq k\eta} \|\mathbf{y}_t - \mathbf{y}_{(k-1)\eta}\|_{L_2} \Bigg\}^2. \quad (65)
\end{aligned}
$$

We need one more result, which provides an upper bound for $\sup_{(k-1)\eta \leq t \leq k\eta} \|\mathbf{y}_t - \mathbf{y}_{(k-1)\eta}\|_{L_2}$. The proof of Lemma B.5 is given in Appendix C.1.

Lemma B.5. *For any* $k = 1, 2, \ldots, K,$

$$
\begin{aligned}
\sup_{(k-1)\eta \leq t \leq k\eta} \left\|\mathbf{y}_t - \mathbf{y}_{(k-1)\eta}\right\|_{L_2} &\leq (\theta(T) + \omega(T)) \int_{(k-1)\eta}^{k\eta} \left[f(T-s) + \frac{1}{2}(g(T-s))^2 L(T-s) \right] \mathrm{d}s \\
&\quad + (L_1 T + \|\nabla \log p_0(\mathbf{0})\|) \int_{(k-1)\eta}^{k\eta} \frac{1}{2}(g(T-s))^2 \mathrm{d}s.
\end{aligned}
$$

where we recall from (31)-(32) *that*

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

$$
\omega(T) := \sup_{0 \leq t \leq T} \left(e^{-2\int_0^t f(s)\mathrm{d}s} \|\mathbf{x}_0\|_{L_2}^2 + d\int_0^t e^{-2\int_s^t f(v)\mathrm{d}v}(g(s))^2\mathrm{d}s \right)^{1/2}.
$$

By applying Lemma B.5, we conclude from (65) that

$$
\begin{aligned}
&\|\mathbf{y}_{k\eta} - \hat{\mathbf{u}}_{k\eta}\|_{L_2}^2 \\
&\leq \Bigg\{ \left(1 - \int_{(k-1)\eta}^{k\eta} \delta_k(T-t)\mathrm{d}t + \frac{L_1}{2}\eta \int_{(k-1)\eta}^{k\eta} (g(T-t))^2\mathrm{d}t \right) e^{\int_{(k-1)\eta}^{k\eta} f(T-t)\mathrm{d}t} \left\|\mathbf{y}_{(k-1)\eta} - \hat{\mathbf{u}}_{(k-1)\eta}\right\|_{L_2} \\
&\quad + \left(\frac{L_1}{2}\eta\left(1 + \|\mathbf{x}_0\|_{L_2} + \omega(T)\right) + \frac{M}{2} \right) \int_{(k-1)\eta}^{k\eta} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s}(g(T-t))^2\mathrm{d}t \\
&\quad + \sqrt{\eta}\nu_{k,\eta} \left(\int_{(k-1)\eta}^{k\eta} \left[\frac{1}{2}e^{\int_t^{k\eta} f(T-s)\mathrm{d}s}(g(T-t))^2 L(T-t)\right]^2 \mathrm{d}t \right)^{1/2} \Bigg\}^2,
\end{aligned}
$$

where $\nu_{k,\eta}$ is defined in (33). The proof of Proposition B.2 is hence complete. $\square$

B.2 Proof of Proposition 3.12

Proof of Proposition 3.12. First of all, we have

RHS of (12)

$$\geq e^{-\int_0^{K\eta} \mu(t)\mathrm{d}t}\|\mathbf{x}_0\|_{L_2} + \sum_{k=1}^{K}\prod_{j=k+1}^{K}\left(1 - \int_{(j-1)\eta}^{j\eta}\delta_j(T-t)\mathrm{d}t\right)$$

$$\cdot \sqrt{\eta}\nu_{k,\eta}\left(\int_{(k-1)\eta}^{k\eta}\left[\frac{1}{2}(g(T-t))^2 L(T-t)\right]^2 \mathrm{d}t\right)^{1/2}$$

$$\geq e^{-\int_0^{K\eta}\mu(t)\mathrm{d}t}\|\mathbf{x}_0\|_{L_2} + \sum_{k=1}^{K}\left(1 - \eta\max_{1\leq j\leq K}\max_{(j-1)\eta\leq t\leq j\eta}\delta_j(T-t)\right)^{K-k}\cdot\eta\nu_{k,\eta}\min_{0\leq t\leq T}\left(\frac{1}{2}(g(t))^2 L(t)\right)$$

$$\geq e^{-\int_0^{K\eta}\mu(t)\mathrm{d}t}\|\mathbf{x}_0\|_{L_2} + \sum_{k=1}^{K}\left(1 - \max_{0\leq t\leq T}\mu(T-t)\right)^{K-k}\cdot\eta\nu_{k,\eta}\min_{0\leq t\leq T}\left(\frac{1}{2}(g(t))^2 L(t)\right)$$

$$= e^{-\int_0^{K\eta}\mu(t)\mathrm{d}t}\|\mathbf{x}_0\|_{L_2} + \frac{1 - (1 - \eta\max_{0\leq t\leq T}\mu(t))^K}{\max_{0\leq t\leq T}\mu(t)}\cdot\nu_{k,\eta}\min_{0\leq t\leq T}\left(\frac{1}{2}(g(t))^2 L(t)\right)$$

$$\geq e^{-\int_0^{K\eta}\mu(t)\mathrm{d}t}\|\mathbf{x}_0\|_{L_2} + \frac{1 - e^{-K\eta\max_{0\leq t\leq T}\mu(t)}}{\max_{0\leq t\leq T}\mu(t)}\cdot\nu_{k,\eta}\min_{0\leq t\leq T}\left(\frac{1}{2}(g(t))^2 L(t)\right),$$

where $\mu(t)$ is given in (24), $\delta_j(T-t)$ is defined in (30) and 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.

Next, by the definition of $\nu_{k,\eta}$ in (33), we have

$$\nu_{k,\eta} \geq \eta\sqrt{d}\left(\int_0^T e^{-2\int_s^T f(v)\mathrm{d}v}(g(s))^2\mathrm{d}s\right)^{1/2}\min_{0\leq t\leq T}(g(t))^2 L(t).$$

Therefore, in order for RHS of (12) $\leq \epsilon$, we must have

$$e^{-\int_0^{K\eta}\mu(t)\mathrm{d}t}\|\mathbf{x}_0\|_{L_2} \leq \epsilon,$$

which implies that $K\eta \to \infty$ as $\epsilon \to 0$ and in particular

$$T = K\eta = \Omega(1), \tag{66}$$

(since under our assumptions on f and g, $\mu(t)$ which is defined in (24) is positive and continuous so that $\int_0^t \mu(s)\mathrm{d}s$ is finite for any $t \in (0,\infty)$ and strictly increasing from 0 to ∞ as t increases from 0 to ∞), and we also need

$$\frac{1 - e^{-K\eta\max_{0\leq t\leq T}\mu(t)}}{\max_{0\leq t\leq T}\mu(t)}\cdot\eta\sqrt{d}\left(\int_0^T e^{-2\int_s^T f(v)\mathrm{d}v}(g(s))^2\mathrm{d}s\right)^{1/2}\cdot\frac{1}{2}\left(\min_{0\leq t\leq T}(g(t))^2 L(t)\right)^2 \leq \epsilon. \tag{67}$$

Note that $\max_{0\leq t\leq T}\mu(t) = \max_{0\leq t\leq K\eta}\mu(t)$. Together with $e^{-\int_0^{K\eta}\mu(t)\mathrm{d}t} = \mathcal{O}(\epsilon/\sqrt{d})$ (since $\|\mathbf{x}_0\|_{L_2} \leq \mathcal{O}(\sqrt{d})$ from (77)) and the assumption that $\max_{0\leq s\leq t}\mu(s) \leq c_1\left(\int_0^t \mu(s)\mathrm{d}s\right)^\rho + c_2$ uniformly in t for some $c_1, c_2, \rho > 0$, it is easy to see that $\max_{0\leq t\leq K\eta}\mu(t) = \mathcal{O}\left(\left(\log\left(\sqrt{d}/\epsilon\right)\right)^\rho\right)$. Moreover, since we assumed $\min_{t\geq 0}(g(t))^2 L(t) > 0$, we have $\mu(t) > 0$ for any t. Together with $T = K\eta = \Omega(1)$ from (66), we have $\max_{0\leq t\leq T}\mu(t) \geq \Omega(1)$. Since $K\eta \to \infty$ as $\epsilon \to 0$, we have $1 - e^{-K\eta\max_{0\leq t\leq T}\mu(t)} = \Omega(1)$. Therefore, it follows from (67) that $\eta = \widetilde{\mathcal{O}}\left(\frac{\epsilon}{\sqrt{d}}\right)$, where $\widetilde{\mathcal{O}}$ ignores the logarithmic dependence on ϵ and d and we used the assumption $\min_{t\geq 0}(g(t))^2 L(t) > 0$ and $\liminf_{T\to\infty}\int_0^T e^{-2\int_s^T f(v)\mathrm{d}v}(g(s))^2\mathrm{d}s > 0$. Hence, we conclude that we have the following lower bound for the complexity: $K = \widetilde{\Omega}\left(\frac{\sqrt{d}}{\epsilon}\right)$, where $\widetilde{\Omega}$ ignores the logarithmic dependence on ϵ and d. This completes the proof. $\square$

C Additional Technical Proofs

C.1 Proof of Lemma B.5

Proof of Lemma B.5. We can compute that for any $(k-1)\eta \leq t \leq k\eta$,

$$\mathbf{y}_t - \mathbf{y}_{(k-1)\eta} = \int_{(k-1)\eta}^{t} \left[f(T-s)\mathbf{y}_s + \frac{1}{2}(g(T-s))^2 \nabla \log p_{T-s}(\mathbf{y}_s) \right] \mathrm{d}s,$$

and moreover

$$\tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_{(k-1)\eta} = \int_{(k-1)\eta}^{t} \left[f(T-s)\tilde{\mathbf{x}}_s + \frac{1}{2}(g(T-s))^2 \nabla \log p_{T-s}(\tilde{\mathbf{x}}_s) \right] \mathrm{d}s, \tag{68}$$

so that

$$\mathbf{y}_t - \mathbf{y}_{(k-1)\eta} = \tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_{(k-1)\eta}$$
$$+ \int_{(k-1)\eta}^{t} \left[f(T-s)(\mathbf{y}_s - \tilde{\mathbf{x}}_s) + \frac{1}{2}(g(T-s))^2 \left(\nabla \log p_{T-s}(\mathbf{y}_s) - \nabla \log p_{T-s}(\tilde{\mathbf{x}}_s) \right) \right] \mathrm{d}s,$$

and therefore

$$\left\| \mathbf{y}_t - \mathbf{y}_{(k-1)\eta} \right\|_{L_2}$$
$$\leq \left\| \tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_{(k-1)\eta} \right\|_{L_2} + \int_{(k-1)\eta}^{t} \left[f(T-s) + \frac{1}{2}(g(T-s))^2 L(T-s) \right] \|\mathbf{y}_s - \tilde{\mathbf{x}}_s\|_{L_2} \mathrm{d}s.$$

We obtained in the proof of Proposition B.1 that

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

and moreover, from the proof of Proposition B.1, we have $\|\mathbf{y}_0 - \tilde{\mathbf{x}}_0\|_{L_2} \leq e^{-\int_0^T f(s)\mathrm{d}s} \|\mathbf{x}_0\|_{L_2}$ so that

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

Therefore, we have

$$\left\| \mathbf{y}_t - \mathbf{y}_{(k-1)\eta} \right\|_{L_2}$$
$$\leq \left\| \tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_{(k-1)\eta} \right\|_{L_2} + \theta(T) \int_{(k-1)\eta}^{k\eta} \left[f(T-s) + \frac{1}{2}(g(T-s))^2 L(T-s) \right] \mathrm{d}s, \tag{69}$$

where $\theta(T)$ bounds $\sup_{0 \leq t \leq T} \|\mathbf{y}_t - \tilde{\mathbf{x}}_t\|_{L_2}$ and it is given in (31).

Next, it follows from (68) that

$$\left\| \tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_{(k-1)\eta} \right\|_{L_2} \leq \int_{(k-1)\eta}^{t} \left[f(T-s)\|\tilde{\mathbf{x}}_s\|_{L_2} + \frac{1}{2}(g(T-s))^2 \|\nabla \log p_{T-s}(\tilde{\mathbf{x}}_s)\|_{L_2} \right] \mathrm{d}s, \tag{70}$$

and we have

$$\|\nabla \log p_{T-s}(\tilde{\mathbf{x}}_s)\|_{L_2} \leq \|\nabla \log p_{T-s}(\tilde{\mathbf{x}}_s) - \nabla \log p_{T-s}(\mathbf{0})\|_{L_2} + \|\nabla \log p_{T-s}(\mathbf{0})\|_{L_2}$$
$$\leq L(T-s) \|\tilde{\mathbf{x}}_s\|_{L_2} + \|\nabla \log p_{T-s}(\mathbf{0})\|$$
$$\leq L(T-s) \|\tilde{\mathbf{x}}_s\|_{L_2} + L_1 T + \|\nabla \log p_0(\mathbf{0})\|, \tag{71}$$

where we applied Assumption 3.2 to obtain the last inequality above.

Therefore, we have

$$\left\| \tilde{\mathbf{x}}_t - \tilde{\mathbf{x}}_{(k-1)\eta} \right\|_{L_2} \leq \omega(T) \int_{(k-1)\eta}^{t} \left[f(T-s) + \frac{1}{2}(g(T-s))^2 L(t-s) \right] \mathrm{d}s$$
$$+ (L_1 T + \|\nabla \log p_0(\mathbf{0})\|) \int_{(k-1)\eta}^{t} \frac{1}{2}(g(T-s))^2 \mathrm{d}s, \tag{72}$$

where we recall that $\tilde{\mathbf{x}}_s$ has the same distribution as $\mathbf{x}_{T-s}$ and we also recall from (63) that $\omega(T) = \sup_{0 \le t \le T} \|\mathbf{x}_t\|_{L_2}$ with an explicit formula given in (32) (see the derivation that leads to (63) in the proof of Proposition B.2) Hence, we conclude that uniformly for $(k-1)\eta \le t \le k\eta$,

$$\left\|\mathbf{y}_t - \mathbf{y}_{(k-1)\eta}\right\|_{L_2} \le (\theta(T) + \omega(T)) \int_{(k-1)\eta}^{k\eta} \left[f(T-s) + \frac{1}{2}(g(T-s))^2 L(T-s) \right] \mathrm{d}s$$

$$+ (L_1 T + \|\nabla \log p_0(\mathbf{0})\|) \int_{(k-1)\eta}^{k\eta} \frac{1}{2}(g(T-s))^2 \mathrm{d}s.$$

This completes the proof. $\qquad\square$

D Derivation of Results in Section 3.3

In this section, we prove the results that are summarized in Table 1 in Section 3.3. We discuss variance exploding SDEs in Appendix D.1, variance preserving SDEs in Appendix D.2.

D.1 Variance-Exploding SDEs

In this section, we consider variance-exploding SDEs with $f(t) \equiv 0$ in the forward process (1). We can immediately obtain the following corollary of Theorem 3.4.

Corollary D.1. *Assume that Assumptions 3.1, 3.2, and 3.3 hold and $\eta \le \bar{\eta}$, where $\bar{\eta} > 0$ is defined in (27). Then, we have*

$$\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0) \le e^{-\int_0^{K\eta} \mu(t)\mathrm{d}t} \|\mathbf{x}_0\|_{L_2}$$

$$+ \sum_{k=1}^{K} \prod_{j=k+1}^{K} \left(1 - \int_{(j-1)\eta}^{j\eta} \delta(T-t)\mathrm{d}t + \frac{L_1}{2}\eta \int_{(j-1)\eta}^{j\eta} (g(T-t))^2 \mathrm{d}t \right)$$

$$\cdot \left(\frac{L_1}{2}\eta \left(1 + 2\|\mathbf{x}_0\|_{L_2} + \sqrt{d} \left(\int_0^T (g(t))^2 \mathrm{d}t \right)^{1/2} \right) \int_{(k-1)\eta}^{k\eta} (g(T-t))^2 \mathrm{d}t \right.$$

$$\left. + \frac{M}{2} \int_{(k-1)\eta}^{k\eta} (g(T-t))^2 \mathrm{d}t + \sqrt{\eta}\nu_{k,\eta} \left(\int_{(k-1)\eta}^{k\eta} \frac{1}{4}(g(T-t))^4 (L(T-t))^2 \mathrm{d}t \right)^{1/2} \right), \qquad (73)$$

where for any $0 \le t \le T$:

$$\delta(T-t) := \frac{\frac{1}{2}(g(T-t))^2}{\frac{1}{m_0} + \int_0^{T-t}(g(s))^2 \mathrm{d}s} - \frac{\eta}{4}(g(T-t))^4 (L(T-t))^2, \qquad (74)$$

where $L(T-t) := \min\left(\left(\int_0^{T-t}(g(s))^2 \mathrm{d}s \right)^{-1}, L_0 \right)$ and

$$\nu_{k,\eta} := \left(2\|\mathbf{x}_0\|_{L_2} + \sqrt{d} \left(\int_0^T (g(t))^2 \mathrm{d}t \right)^{1/2} \right) \int_{(k-1)\eta}^{k\eta} \frac{1}{2}(g(T-s))^2 L(T-s) \mathrm{d}s$$

$$+ (L_1 T + \|\nabla \log p_0(\mathbf{0})\|) \int_{(k-1)\eta}^{k\eta} \frac{1}{2}(g(T-s))^2 \mathrm{d}s, \qquad (75)$$

where $\mu(t)$ is defined as:

$$\mu(t) := \frac{\frac{1}{2}(g(t))^2}{\frac{1}{m_0} + \int_0^t (g(s))^2 \mathrm{d}s}. \qquad (76)$$

The term $\|\mathbf{x}_0\|_{L_2}$ in Corollary D.1 has square root dependence on the dimension d. Indeed, by Lemma 11 in Gürbüzbalaban et al. (2021), we have

$$\|\mathbf{x}_0\|_{L_2} \le \sqrt{2d/m_0} + \|\mathbf{x}_*\|, \qquad (77)$$

where $\mathbf{x}_*$ is the unique minimizer of $-\log p_0$.

In the next few sections, we consider special functions g in Corollary D.1 and derive the corresponding results in Table 1.

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

When $g(t) = ae^{bt}$ for some $a, b > 0$, we can obtain the following result from Corollary D.1.

Corollary D.2 (Restatement of Corollary 3.8). *Let $g(t) = ae^{bt}$ for some $a, b > 0$. Then, we have $\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0) \leq \mathcal{O}(\epsilon)$ after $K = \mathcal{O}\left(\frac{d^{3/2}\log(d/\epsilon)}{\epsilon^3}\right)$ iterations provided that $M \leq \frac{\epsilon^2}{\sqrt{d}}$ and $\eta \leq \frac{\epsilon^3}{d^{3/2}}$.*

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

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

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 $\mu(t)$ in (76), we can compute that

$$\mu(t) = \frac{\frac{1}{2}m_0(g(t))^2}{1 + m_0 \int_0^t (g(s))^2 \mathrm{d}s} = \frac{\frac{1}{2}m_0 a^2 e^{2bt}}{1 + m_0 \frac{a^2}{2b}(e^{2bt} - 1)}. \tag{78}$$

This implies that

$$\int_0^t \mu(s)\mathrm{d}s = \frac{1}{2}\int_0^t \frac{2bm_0 a^2 e^{2bs}\mathrm{d}s}{2b - m_0 a^2 + m_0 a^2 e^{2bs}} = \frac{1}{2}\log\left(\frac{2b - m_0 a^2 + m_0 a^2 e^{2bt}}{2b}\right). \tag{79}$$

By letting $t = T = K\eta$ in (79) and using (77), we obtain

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

Moreover,

$$\nu_{k,\eta} \leq \left(2\sqrt{2d/m_0} + 2\|\mathbf{x}_*\| + \sqrt{d}\frac{a}{\sqrt{2b}}\left(e^{2bT} - 1\right)^{1/2}\right)\max\left(4b, \frac{L_0 a^2}{4b^2}\right)\frac{\eta}{2}$$
$$+ (L_1 T + \|\nabla \log p_0(\mathbf{0})\|)\frac{a^2}{4b}\left(e^{2b(T-(k-1)\eta)} - e^{2b(T-k\eta)}\right), \tag{80}$$

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

$$0 \leq 1 - \int_{(j-1)\eta}^{j\eta} \delta(T - t)\mathrm{d}t + \frac{L_1}{2}\eta\int_{(j-1)\eta}^{j\eta}(g(T-t))^2 \mathrm{d}t \leq 1,$$

where $\delta(\cdot)$ is defined in (74), provided that the condition $\eta \leq \bar{\eta}$ (where $\bar{\eta}$ is defined in (27)) holds, that is:

$$\eta \leq \min\left\{\min_{0 \leq t \leq T}\left\{\frac{\frac{1}{2}(g(t))^2}{\frac{1}{m_0} + \int_0^t (g(s))^2 \mathrm{d}s}\right\}, \min_{0 \leq t \leq T}\left\{\frac{\frac{1}{m_0} + \int_0^t (g(s))^2 \mathrm{d}s}{\frac{1}{2}(g(t))^2}\right\}\right\}, \tag{81}$$

which holds provided that

$$\eta \leq \min\left\{\min_{0 \leq t \leq T}\left\{\frac{\frac{\frac{1}{2}a^2 e^{2bt}}{\frac{1}{m_0} + \frac{a^2}{2b}(e^{2bt}-1)}}{\frac{1}{4}\max\left(16b^2, \frac{L_0^2 a^4}{16b^4}\right) + \frac{L_1}{2}a^2 e^{2bt}}\right\}, \min_{0 \leq t \leq T}\left\{\frac{\frac{1}{m_0} + \frac{a^2}{2b}(e^{2bt} - 1)}{\frac{1}{2}a^2 e^{2bt}}\right\}\right\}. \tag{82}$$

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

$$\prod_{j=k+1}^{K} \left(1 - \int_{(j-1)\eta}^{j\eta} \delta(T-t)\mathrm{d}t + \frac{L_1}{2}\eta \int_{(j-1)\eta}^{j\eta} (g(T-t))^2\mathrm{d}t \right)$$

$$\leq \prod_{j=k+1}^{K} e^{-\int_{(j-1)\eta}^{j\eta} \delta(T-t)\mathrm{d}t + \frac{L_1}{2}\eta \int_{(j-1)\eta}^{j\eta} (g(T-t))^2\mathrm{d}t} = e^{-\int_{k\eta}^{K\eta} \delta(T-t)\mathrm{d}t + \frac{L_1}{2}\eta \int_{k\eta}^{K\eta} (g(T-t))^2\mathrm{d}t},$$

where $\delta(\cdot)$ is defined in (74). Moreover,

$$\int_{k\eta}^{K\eta} \delta(T-t)\mathrm{d}t \geq \int_{k\eta}^{K\eta} \frac{\frac{1}{2}m_0 a^2 e^{2b(T-t)}\mathrm{d}t}{1 + m_0 \frac{a^2}{2b}(e^{2b(T-t)} - 1)} - \frac{1}{4}(K-k)\eta^2 \max\left(16b^2, \frac{L_0^2 a^4}{16b^4}\right)$$

$$= \frac{1}{2}\log\left(\frac{2b - m_0 a^2 + m_0 a^2 e^{2b(T-k\eta)}}{2b - m_0 a^2 + m_0 a^2 e^{2b(T-K\eta)}} \right) - \frac{1}{4}(K-k)\eta^2 \max\left(16b^2, \frac{L_0^2 a^4}{16b^4}\right),$$

and

$$\frac{L_1}{2}\eta \int_{k\eta}^{K\eta} (g(T-t))^2\mathrm{d}t = \frac{L_1}{2}\eta \frac{a^2}{2b}\left(e^{2b(K-k)\eta} - 1 \right).$$

By applying Corollary D.1 with $T = K\eta$, we conclude that

$$\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0)$$

$$\leq \frac{\sqrt{2b}\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\| \right)}{\sqrt{2b - m_0 a^2 + m_0 a^2 e^{2bK\eta}}}$$

$$+ \sum_{k=1}^{K} \frac{\sqrt{2b}}{\sqrt{2b - m_0 a^2 + m_0 a^2 e^{2b(K-k)\eta}}} e^{(K-k)\eta^2 \frac{1}{4}\max\left(16b^2, \frac{L_0^2 a^4}{16b^4}\right) + \frac{L_1}{2}\eta \frac{a^2}{2b}(e^{2b(K-k)\eta} - 1)}$$

$$\cdot \left(\left(\frac{M}{2} + \frac{L_1}{2}\eta\left(1 + 2\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\| \right) + \sqrt{d}\frac{a}{\sqrt{2b}}(e^{2bK\eta} - 1)^{1/2} \right) \right) \right)$$

$$\cdot \frac{a^2}{2b}\left(e^{2b(K-k+1)\eta} - e^{2b(K-k)\eta} \right)$$

$$+ \frac{\eta}{2}\max\left(4b, \frac{L_0 a^2}{4b^2} \right) \cdot \left(\left(2\sqrt{2d/m_0} + 2\|\mathbf{x}_*\| + \sqrt{d}\frac{a}{\sqrt{2b}}\left(e^{2bT} - 1 \right)^{1/2} \right) \max\left(4b, \frac{L_0 a^2}{4b^2} \right)\frac{\eta}{2}$$

$$+ (L_1 T + \|\nabla\log p_0(\mathbf{0})\|) \frac{a^2}{4b}\left(e^{2b(T-(k-1)\eta)} - e^{2b(T-k\eta)} \right) \right) \Bigg).$$

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

$$
\begin{aligned}
\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0) & \\
\leq & \mathcal{O}\left(\frac{\sqrt{d}}{e^{bK\eta}}\right) + \mathcal{O}\left(e^{\frac{1}{4}K\eta^2 \max\left(16b^2, \frac{L_0^2 a^4}{16b^4}\right) + \frac{L_1}{2}\eta\frac{a^2}{2b}e^{2bK\eta}}\right. \\
& \left. \cdot \sum_{k=1}^{K}\frac{1}{e^{b(K-k)\eta}} \cdot \left(\left(M + L_1\eta\sqrt{d}e^{bK\eta}\right)e^{2b(K-(k-1))\eta}\eta + \eta e^{bK\eta}\eta\sqrt{d}\right.\right. \\
& \left.\left. + \eta^2 e^{2b(K-(k-1))\eta} + e^{2b(K-(k-1))\eta}\eta^2(K\eta)\right)\right) \\
\leq & \mathcal{O}\left(\frac{\sqrt{d}}{e^{bK\eta}}\right) \\
& + \mathcal{O}\left(e^{K\eta^2 \max\left(16b^2, \frac{L_0^2 a^4}{16b^4}\right)} \cdot \left(\left(M + L_1\eta\sqrt{d}e^{bK\eta}\right)e^{bK\eta} + K\eta^2\sqrt{d} + \eta(K\eta)e^{bK\eta}\right)\right) \\
\leq & \mathcal{O}(\epsilon),
\end{aligned}
$$

and (82) holds such that the condition $\eta \leq \bar{\eta}$ (where $\bar{\eta}$ is defined in (27)) holds provided that

$$
K\eta = \frac{\log(\sqrt{d}/\epsilon)}{b}, \qquad M \leq \frac{\epsilon^2}{\sqrt{d}}, \qquad \eta \leq \frac{\epsilon^3}{d^{3/2}},
$$

which implies that $K \geq \mathcal{O}\left(\frac{d^{3/2}\log(d/\epsilon)}{\epsilon^3}\right)$. This completes the proof. $\qquad\square$

D.1.2   Example: $f(t) \equiv 0$ and $g(t) = (b + at)^c$

Karras et al. (2022) considers $f(t) \equiv 0, g(t) = \sqrt{2t}$ with non-uniform discretization time steps, where the time steps are defined according to a polynomial noise schedule. Inspired by Karras et al. (2022), we next consider $g(t) = (b + at)^c$ for some $a, b, c > 0$, we can obtain the following result from Corollary D.1.

Corollary D.3 (Restatement of Corollary 3.9)**.** *Let $g(t) = (b + at)^c$ for some $a, b > 0$, $c \geq 1/2$. Then, we have* $\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0) \leq \mathcal{O}(\epsilon)$ *after* $K = \mathcal{O}\left(\frac{d^{\frac{1}{(2c+1)}+\frac{3}{2}}}{\epsilon^{\frac{2}{2c+1}+3}}\right)$ *iterations provided that* $M \leq \frac{\epsilon^2}{\sqrt{d}}$ *and* $\eta \leq \frac{\epsilon^3}{d^{\frac{3}{2}}}$.

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

$$
(g(t))^2 L(t) = \min\left(\frac{(b+at)^c}{\frac{1}{a(2c+1)}((b+at)^{2c+1} - b^{2c+1})}, L_0(b+at)^{2c}\right). \tag{83}
$$

If $t \geq \frac{b}{a}$, then

$$
\frac{(b+at)^c}{\frac{1}{a(2c+1)}((b+at)^{2c+1} - b^{2c+1})} \leq \frac{(b+at)^c}{\frac{1}{a(2c+1)}(1 - \frac{1}{2^{2c+1}})(b+at)^{2c+1}} \leq \frac{a(2c+1)}{(1 - \frac{1}{2^{2c+1}})b}. \tag{84}
$$

If $t \leq \frac{b}{a}$, then

$$
L_0(b+at)^{2c} \leq L_0(2b)^{2c}. \tag{85}
$$

Therefore, it follows from (83), (84) and (85) 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 (76), we have

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

Furthermore,

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

$$
\cdot \frac{1}{2} \max \left(\frac{a(2c+1)}{(1 - \frac{1}{2^{2c+1}})b}, L_0(2b)^{2c} \right) \eta
$$

$$
+ (L_1 T + \|\nabla \log p_0(\mathbf{0})\|) \frac{1}{2} \frac{(b + a(T - (k-1)\eta))^{2c+1} - (b + a(T - k\eta))^{2c+1}}{a(2c+1)}, \tag{86}
$$

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

$$
\delta(t) = \frac{\frac{1}{2}(b+at)^{2c}}{\frac{1}{m_0} + \frac{1}{a(2c+1)}((b+at)^{2c+1} - b^{2c+1})}
$$

$$
- \frac{1}{4}\eta \min \left(\frac{(b+at)^{2c}}{\frac{1}{a^2(2c+1)^2}((b+at)^{2c+1} - b^{2c+1})^2}, L_0^2(b+at)^{4c} \right),
$$

(where $\delta(\cdot)$ is defined in (74)) satisfies

$$
0 \leq 1 - \int_{(j-1)\eta}^{j\eta} \delta(T - t)\mathrm{d}t + \frac{L_1}{2}\eta \int_{(j-1)\eta}^{j\eta} (g(T-t))^2 \mathrm{d}t \leq 1,
$$

provided that the condition $\eta \leq \bar{\eta}$ (where $\bar{\eta}$ is defined in (27)) holds.

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

$$
\prod_{j=k+1}^{K} \left(1 - \int_{(j-1)\eta}^{j\eta} \delta(T-t)\mathrm{d}t + \frac{L_1}{2}\eta \int_{(j-1)\eta}^{j\eta} (g(T-t))^2 \mathrm{d}t \right)
$$

$$
\leq \prod_{j=k+1}^{K} e^{- \int_{(j-1)\eta}^{j\eta} \delta(T-t)\mathrm{d}t + \frac{L_1}{2}\eta \int_{(j-1)\eta}^{j\eta} (g(T-t))^2 \mathrm{d}t} = e^{- \int_{k\eta}^{K\eta} \delta(T-t)\mathrm{d}t + \frac{L_1}{2}\eta \int_{k\eta}^{K\eta} (g(T-t))^2 \mathrm{d}t},
$$

where $\delta(\cdot)$ is defined in (74). Moreover,

$$
\int_{k\eta}^{K\eta} \delta(T - t)\mathrm{d}t \geq \frac{1}{2} \log \left(1 + \frac{m_0}{a(2c+1)} \left((b + a(K-k)\eta)^{2c+1} - b^{2c+1} \right) \right)
$$

$$
- \frac{1}{4}(K-k)\eta^2 \max \left(\frac{a^2(2c+1)^2}{(1 - \frac{1}{2^{2c+1}})^2 b^2}, L_0^2(2b)^{4c} \right),
$$

and we can compute that

$$
\frac{1}{2}L_1\eta \int_{k\eta}^{K\eta} (g(T-t))^2 \mathrm{d}t = \frac{1}{2} \frac{L_1\eta}{a(2c+1)} \left((b + a(K-k)\eta)^{2c+1} - b^{2c+1} \right).
$$

By applying Corollary D.1 with $T = K\eta$ and (77), we conclude that

$$\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0)$$

$$\leq \frac{\sqrt{2d/m_0} + \|\mathbf{x}_*\|}{\sqrt{1 + \frac{m_0}{a(2c+1)}((b + aK\eta)^{2c+1} - b^{2c+1})}}$$

$$+ \sum_{k=1}^K \frac{e^{\frac{1}{4}(K-k)\eta^2 \max\left(\frac{a^2(2c+1)^2}{(1-\frac{1}{2^{2c+1}})^2 b^2}, L_0^2(2b)^{4c}\right) + \frac{1}{2}\frac{L_1\eta}{a(2c+1)}((b + a(K-k)\eta)^{2c+1} - b^{2c+1})}}{\sqrt{1 + \frac{m_0}{a(2c+1)}((b + a(K-k)\eta)^{2c+1} - b^{2c+1})}}$$

$$\cdot \left(\left(\frac{L_1}{2}\eta\left(1 + 2\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right) + \sqrt{d}\left(\frac{(b + aK\eta)^{2c+1} - b^{2c+1}}{a(2c+1)}\right)^{1/2}\right) + \frac{M}{2}\right)\right.$$

$$\cdot \frac{(b + a(K\eta - (k-1)\eta))^{2c+1} - (b + a(K\eta - k\eta))^{2c+1}}{a(2c+1)}$$

$$+ \frac{1}{2}\eta \max\left(\frac{a(2c+1)}{(1 - \frac{1}{2^{2c+1}})b}, L_0(2b)^{2c}\right)$$

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

$$\cdot \frac{1}{2}\max\left(\frac{a(2c+1)}{(1 - \frac{1}{2^{2c+1}})b}, L_0(2b)^{2c}\right)\eta$$

$$\left.\left.+ (L_1 K\eta + \|\nabla \log p_0(\mathbf{0})\|)\frac{1}{2}\frac{(b + a(K\eta - (k-1)\eta))^{2c+1} - (b + a(K\eta - k\eta))^{2c+1}}{a(2c+1)}\right)\right).$$

This implies that

$$\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0)$$

$$\leq \mathcal{O}\left(\frac{\sqrt{d}}{(K\eta)^{\frac{2c+1}{2}}} + e^{\mathcal{O}((K\eta)\eta + (K\eta)^{2c+1}\eta)}\sum_{k=1}^K \frac{1}{((K-k)\eta)^{\frac{2c+1}{2}}}\right.$$

$$\left.\cdot \left(\left(\sqrt{d}(K\eta)^{\frac{2c+1}{2}}L_1\eta + M\right)((K-k)\eta)^{2c}\eta + \eta\left(\eta\sqrt{d}(K\eta)^{\frac{2c+1}{2}} + (K\eta)\eta((K-k)\eta)^{2c}\right)\right)\right)$$

$$\leq \mathcal{O}\left(\frac{\sqrt{d}}{(K\eta)^{\frac{2c+1}{2}}} + e^{\mathcal{O}((K\eta)\eta + (K\eta)^{2c+1}\eta)}\right.$$

$$\left.\cdot \left(\left(\sqrt{d}(K\eta)^{\frac{2c+1}{2}}L_1\eta + M\right)(K\eta)^{c+\frac{1}{2}} + \eta\left(K\eta\sqrt{d} + (K\eta)(K\eta)^{c+\frac{1}{2}}\right)\right)\right)$$

$$\leq \mathcal{O}(\epsilon),$$

and the condition $\eta \leq \bar{\eta}$ (where $\bar{\eta}$ is defined in (27)) holds provided that

$$K\eta = \frac{d^{\frac{1}{(2c+1)}}}{\epsilon^{\frac{2}{2c+1}}}, \qquad M \leq \frac{\epsilon^2}{\sqrt{d}}, \qquad \eta \leq \frac{\epsilon^3}{d^{\frac{3}{2}}},$$

so that $K \geq \mathcal{O}\left(\frac{d^{\frac{1}{(2c+1)} + \frac{3}{2}}}{\epsilon^{\frac{2}{2c+1} + 3}}\right)$. This completes the proof. $\qquad\square$

D.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), where $\beta(t)$ is often chosen as some non-decreasing function in practice. We can obtain the following corollary of Theorem 3.4.

Corollary D.4. *Under the assumptions of Theorem 3.4, we have*

$$
\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0) \leq \frac{\|\mathbf{x}_0\|_{L_2}}{\sqrt{m_0 e^{\int_0^{K\eta} \beta(s)\mathrm{d}s} + 1 - m_0}}
$$

$$
+ \sum_{k=1}^{K} \frac{e^{\int_{k\eta}^{K\eta} \frac{1}{2}\beta(K\eta - t)\mathrm{d}t + \int_{k\eta}^{K\eta} \eta \max(1, L_0^2)(\beta(K\eta - t))^2 \mathrm{d}t + \frac{1}{2}L_1\eta \int_{k\eta}^{K\eta} \beta(K\eta - t)\mathrm{d}t}}{\left(m_0 e^{\int_0^{(K-k)\eta} \beta(s)\mathrm{d}s} + 1 - m_0\right)^{\frac{1}{2}} e^{-\frac{1}{2}\eta \max_{0 \leq t \leq T} \beta(t)}}
$$

$$
\cdot \left(\left(\frac{L_1}{2}\eta\left(1 + \|\mathbf{x}_0\|_{L_2} + \left(\|\mathbf{x}_0\|_{L_2}^2 + d\right)^{1/2}\right) + \frac{M}{2}\right) 2\left(e^{\int_{(k-1)\eta}^{k\eta} \frac{1}{2}\beta(T-v)\mathrm{d}v} - 1\right)\right.
$$

$$
+ \sqrt{\eta} \max(1, L_0) \max_{(k-1)\eta \leq t \leq k\eta} \beta(K\eta - t)\left(e^{\int_{(k-1)\eta}^{k\eta} \beta(T-v)\mathrm{d}v} - 1\right)^{1/2}
$$

$$
\cdot \left(\left(\|\mathbf{x}_0\|_{L_2} + \left(\|\mathbf{x}_0\|_{L_2}^2 + d\right)^{1/2}\right)\left(\frac{1}{2} + \max(1, L_0)\right) \int_{T-k\eta}^{T-(k-1)\eta} \beta(s)\mathrm{d}s\right.
$$

$$
\left.\left. + (L_1 T + \|\nabla \log p_0(\mathbf{0})\|) \int_{T-k\eta}^{T-(k-1)\eta} \frac{1}{2}\beta(s)\mathrm{d}s\right)\right). \tag{87}
$$

Proof. We apply Theorem 3.4 applied to the variance-preserving SDE ($f(t) = \frac{1}{2}\beta(t)$ and $g(t) = \sqrt{\beta(t)}$). First, we can compute that

$$
L(T - t) = \min\left(\left(\int_0^{T-t} e^{-2\int_s^{T-t} f(v)\mathrm{d}v}(g(s))^2\mathrm{d}s\right)^{-1}, \left(e^{\int_0^{T-t} f(s)\mathrm{d}s}\right)^2 L_0\right)
$$

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

If $e^{\int_0^{T-t} \beta(s)\mathrm{d}s} \geq 2$, then $\frac{1}{1 - e^{-\int_0^{T-t} \beta(s)\mathrm{d}s}} \leq 2$ and otherwise $e^{\int_0^{T-t} \beta(s)\mathrm{d}s} L_0 \leq 2L_0$. Therefore, for any $0 \leq t \leq T$,

$$
L(T - t) \leq 2\max(1, L_0).
$$

By applying Theorem 3.4, we have

$$
\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0) \leq e^{-\int_0^{K\eta} \mu(t)\mathrm{d}t}\|\mathbf{x}_0\|_{L_2}
$$

$$
+ \sum_{k=1}^{K} \prod_{j=k+1}^{K}\left(1 - \int_{(j-1)\eta}^{j\eta} \delta_j(T-t)\mathrm{d}t + \frac{L_1}{2}\eta \int_{(j-1)\eta}^{j\eta} (g(T-t))^2\mathrm{d}t\right)
$$

$$
\cdot e^{\int_{k\eta}^{K\eta} f(T-t)\mathrm{d}t}\left(\frac{L_1}{2}\eta\left(1 + \|\mathbf{x}_0\|_{L_2} + \omega(T)\right) \int_{(k-1)\eta}^{k\eta} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s}(g(T-t))^2\mathrm{d}t\right.
$$

$$
+ \frac{M}{2} \int_{(k-1)\eta}^{k\eta} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s}(g(T-t))^2\mathrm{d}t
$$

$$
\left. + \sqrt{\eta}\nu_{k,\eta}\left(\int_{(k-1)\eta}^{k\eta} \left[\frac{1}{2}e^{\int_t^{k\eta} f(T-s)\mathrm{d}s}(g(T-t))^2 L(T-t)\right]^2 \mathrm{d}t\right)^{1/2}\right),
$$

where

$$
\nu_{k,\eta} := (\theta(T) + \omega(T)) \int_{(k-1)\eta}^{k\eta} \left[f(T-s) + \frac{1}{2}(g(T-s))^2 L(T-s)\right]\mathrm{d}s
$$

$$
+ (L_1 T + \|\nabla \log p_0(\mathbf{0})\|) \int_{(k-1)\eta}^{k\eta} \frac{1}{2}(g(T-s))^2\mathrm{d}s,
$$

where

$$\theta(T) = e^{-\int_0^T \frac{1}{2}\beta(s)\mathrm{d}s}\|\mathbf{x}_0\|_{L_2} \leq \|\mathbf{x}_0\|_{L_2},$$

and

$$\omega(T) = \sup_{0 \leq t \leq T} \left(e^{-2\int_0^t f(s)\mathrm{d}s}\|\mathbf{x}_0\|_{L_2}^2 + d\int_0^t e^{-2\int_s^t f(v)\mathrm{d}v}(g(s))^2\mathrm{d}s\right)^{1/2}$$

$$= \sup_{0 \leq t \leq T} \left(e^{-\int_0^t \beta(s)\mathrm{d}s}\|\mathbf{x}_0\|_{L_2}^2 + d\left(1 - e^{-\int_0^t \beta(s)\mathrm{d}s}\right)\right)^{1/2} \leq \left(\|\mathbf{x}_0\|_{L_2}^2 + d\right)^{1/2}.$$

We can compute that

$$\int_{(k-1)\eta}^{k\eta} e^{\int_t^{k\eta} f(T-s)\mathrm{d}s}(g(T-t))^2\mathrm{d}t = \int_{(k-1)\eta}^{k\eta} e^{\int_s^{k\eta} \frac{1}{2}\beta(T-v)\mathrm{d}v}\beta(T-s)\mathrm{d}s$$

$$= 2\left(e^{\int_{(k-1)\eta}^{k\eta} \frac{1}{2}\beta(T-v)\mathrm{d}v} - 1\right).$$

Furthermore, we have

$$\nu_{k,\eta} \leq \left(\|\mathbf{x}_0\|_{L_2} + \left(\|\mathbf{x}_0\|_{L_2}^2 + d\right)^{1/2}\right)\left(\frac{1}{2} + \max(1, L_0)\right)\int_{(k-1)\eta}^{k\eta} \beta(T-s)\mathrm{d}s$$

$$+ (L_1 T + \|\nabla\log p_0(\mathbf{0})\|)\int_{T-k\eta}^{T-(k-1)\eta} \frac{1}{2}\beta(s)\mathrm{d}s$$

$$= \left(\|\mathbf{x}_0\|_{L_2} + \left(\|\mathbf{x}_0\|_{L_2}^2 + d\right)^{1/2}\right)\left(\frac{1}{2} + \max(1, L_0)\right)\int_{T-k\eta}^{T-(k-1)\eta} \beta(s)\mathrm{d}s$$

$$+ (L_1 T + \|\nabla\log p_0(\mathbf{0})\|)\int_{T-k\eta}^{T-(k-1)\eta} \frac{1}{2}\beta(s)\mathrm{d}s.$$

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

$$\mu(t) = \frac{\frac{1}{2}m_0\beta(t)}{e^{-\int_0^t \beta(s)\mathrm{d}s} + m_0\int_0^t e^{-\int_s^t \beta(v)\mathrm{d}v}\beta(s)\mathrm{d}s} = \frac{\frac{1}{2}m_0\beta(t)}{e^{-\int_0^t \beta(s)\mathrm{d}s} + m_0(1 - e^{-\int_0^t \beta(s)\mathrm{d}s})}. \tag{88}$$

It follows that

$$\int_0^T \mu(t)\mathrm{d}t = \frac{1}{2}\int_0^{\int_0^T \beta(s)\mathrm{d}s} \frac{m_0 dx}{m_0 + (1-m_0)e^{-x}} = \frac{1}{2}\log\left(m_0 e^{\int_0^T \beta(s)\mathrm{d}s} + 1 - m_0\right). \tag{89}$$

Hence, we obtain

$$e^{-\int_0^T \mu(t)\mathrm{d}t}\|\mathbf{x}_0\|_{L_2} = \frac{\|\mathbf{x}_0\|_{L_2}}{\sqrt{m_0 e^{\int_0^T \beta(s)\mathrm{d}s} + 1 - m_0}}. \tag{90}$$

Under the condition $\eta \leq \bar{\eta}$ (where $\bar{\eta}$ is defined in (27)),

$$0 \leq 1 - \int_{(j-1)\eta}^{j\eta} \delta_j(T-t)\mathrm{d}t + \frac{L_1}{2}\eta\int_{(j-1)\eta}^{j\eta}(g(T-t))^2\mathrm{d}t \leq 1.$$

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

$$\prod_{j=k+1}^K \left(1 - \int_{(j-1)\eta}^{j\eta} \delta_j(T-t)\mathrm{d}t + \frac{L_1}{2}\eta\int_{(j-1)\eta}^{j\eta}(g(T-t))^2\mathrm{d}t\right)$$

$$\leq \prod_{j=k+1}^K e^{-\int_{(j-1)\eta}^{j\eta} \delta_j(T-t)\mathrm{d}t + \frac{L_1}{2}\eta\int_{(j-1)\eta}^{j\eta}(g(T-t))^2\mathrm{d}t}$$

$$= e^{-\sum_{j=k+1}^K \int_{(j-1)\eta}^{j\eta} \delta_j(T-t)\mathrm{d}t + \frac{L_1}{2}\eta\int_{k\eta}^{K\eta}(g(T-t))^2\mathrm{d}t},$$

where

$$\sum_{j=k+1}^{K}\int_{(j-1)\eta}^{j\eta}\delta_j(T-t)\mathrm{d}t$$

$$=\sum_{j=k+1}^{K}\int_{(j-1)\eta}^{j\eta}\frac{\frac{1}{2}e^{-\int_{(j-1)\eta}^{t}f(T-s)\mathrm{d}s}(g(T-t))^2}{\frac{1}{m_0}e^{-2\int_0^{T-t}f(s)\mathrm{d}s}+\int_0^{T-t}e^{-2\int_s^{T-t}f(v)\mathrm{d}v}(g(s))^2\mathrm{d}s}\mathrm{d}t$$

$$-\sum_{j=k+1}^{K}\int_{(j-1)\eta}^{j\eta}\frac{\eta}{4}(g(T-t))^4(L(T-t))^2\mathrm{d}t$$

$$=\sum_{j=k+1}^{K}\int_{(j-1)\eta}^{j\eta}\frac{\frac{1}{2}e^{-\frac{1}{2}\int_{(j-1)\eta}^{t}\beta(T-s)\mathrm{d}s}\beta(T-t)}{\frac{1}{m_0}e^{-\int_0^{T-t}\beta(s)\mathrm{d}s}+\int_0^{T-t}e^{-\int_s^{T-t}\beta(v)\mathrm{d}v}\beta(s)\mathrm{d}s}\mathrm{d}t$$

$$-\int_{k\eta}^{K\eta}\frac{\eta}{4}(\beta(T-t))^2(L(T-t))^2\mathrm{d}t$$

$$=\sum_{j=k+1}^{K}\int_{(j-1)\eta}^{j\eta}\frac{\frac{1}{2}e^{-\frac{1}{2}\int_{(j-1)\eta}^{t}\beta(T-s)\mathrm{d}s}\beta(T-t)}{\left(\frac{1}{m_0}-1\right)e^{-\int_0^{T-t}\beta(s)\mathrm{d}s}+1}\mathrm{d}t-\int_{k\eta}^{K\eta}\frac{\eta}{4}(\beta(T-t))^2(L(T-t))^2\mathrm{d}t.$$

We can further compute that

$$\sum_{j=k+1}^{K}\int_{(j-1)\eta}^{j\eta}\frac{\frac{1}{2}e^{-\frac{1}{2}\int_{(j-1)\eta}^{t}\beta(T-s)\mathrm{d}s}\beta(T-t)}{\left(\frac{1}{m_0}-1\right)e^{-\int_0^{T-t}\beta(s)\mathrm{d}s}+1}\mathrm{d}t$$

$$\geq e^{-\frac{1}{2}\eta\max_{0\leq t\leq T}\beta(t)\mathrm{d}t}\int_{k\eta}^{K\eta}\frac{\frac{1}{2}\beta(T-t)}{\left(\frac{1}{m_0}-1\right)e^{-\int_0^{T-t}\beta(s)\mathrm{d}s}+1}\mathrm{d}t$$

$$=\frac{1}{2}e^{-\frac{1}{2}\eta\max_{0\leq t\leq T}\beta(t)\mathrm{d}t}\log\left(m_0e^{\int_0^{(K-k)\eta}\beta(s)\mathrm{d}s}+1-m_0\right).$$

Moreover, we can compute that

$$\left(\int_{(k-1)\eta}^{k\eta}\left[\frac{1}{2}e^{\int_t^{k\eta}f(T-s)\mathrm{d}s}(g(T-t))^2L(T-t)\right]^2\mathrm{d}t\right)^{1/2}$$

$$\leq\max(1,L_0)\left(\int_{(k-1)\eta}^{k\eta}e^{\int_t^{k\eta}\beta(T-v)\mathrm{d}v}(\beta(K\eta-t))^2\mathrm{d}t\right)^{1/2}$$

$$\leq\max(1,L_0)\max_{(k-1)\eta\leq t\leq k\eta}\beta(K\eta-t)\left(\int_{(k-1)\eta}^{k\eta}e^{\int_t^{k\eta}\beta(T-v)\mathrm{d}v}\beta(K\eta-t)\mathrm{d}t\right)^{1/2}$$

$$=\max(1,L_0)\max_{(k-1)\eta\leq t\leq k\eta}\beta(K\eta-t)\left(e^{\int_{(k-1)\eta}^{k\eta}\beta(T-v)\mathrm{d}v}-1\right)^{1/2}.$$

By using $T = K\eta$, we conclude that

$$
\begin{aligned}
\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0) \leq {} & \frac{\|\mathbf{x}_0\|_{L_2}}{\sqrt{m_0 e^{\int_0^{K\eta} \beta(s)\mathrm{d}s} + 1 - m_0}} \\
& + \sum_{k=1}^{K} \frac{e^{\int_{k\eta}^{K\eta} \frac{1}{2}\beta(K\eta-t)\mathrm{d}t + \int_{k\eta}^{K\eta} \eta\max(1, L_0^2)(\beta(K\eta-t))^2\mathrm{d}t + \frac{1}{2}L_1\eta\int_{k\eta}^{K\eta}\beta(K\eta-t)\mathrm{d}t}}{\left(m_0 e^{\int_0^{(K-k)\eta} \beta(s)\mathrm{d}s} + 1 - m_0\right)^{\frac{1}{2}e^{-\frac{1}{2}\eta\max_{0\leq t\leq T}\beta(t)}}} \\
& \cdot \left(\left(\frac{L_1}{2}\eta\left(1 + \|\mathbf{x}_0\|_{L_2} + \left(\|\mathbf{x}_0\|_{L_2}^2 + d\right)^{1/2}\right) + \frac{M}{2}\right) 2\left(e^{\int_{(k-1)\eta}^{k\eta} \frac{1}{2}\beta(T-v)\mathrm{d}v} - 1\right)\right. \\
& + \sqrt{\eta}\max(1, L_0) \max_{(k-1)\eta\leq t\leq k\eta} \beta(K\eta - t)\left(e^{\int_{(k-1)\eta}^{k\eta} \beta(T-v)\mathrm{d}v} - 1\right)^{1/2} \\
& \cdot \left(\left(\|\mathbf{x}_0\|_{L_2} + \left(\|\mathbf{x}_0\|_{L_2}^2 + d\right)^{1/2}\right)\left(\frac{1}{2} + \max(1, L_0)\right)\int_{T-k\eta}^{T-(k-1)\eta}\beta(s)\mathrm{d}s\right. \\
& \left.\left. + (L_1 T + \|\nabla\log p_0(\mathbf{0})\|)\int_{T-k\eta}^{T-(k-1)\eta}\frac{1}{2}\beta(s)\mathrm{d}s\right)\right). \quad (91)
\end{aligned}
$$

This completes the proof. $\qquad\square$

D.2.1   Example: $\beta(t) \equiv b$

We consider the special case $\beta(t) \equiv b$ for some $b > 0$. This includes the special case $\beta(t) \equiv 2$ that is studied in Chen et al. (2023) Chen et al. (2023c).

Corollary D.5 (Restatement of Corollary 3.10). *Assume $\beta(t) \equiv b$. Then, we have $\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0) \leq \mathcal{O}(\epsilon)$ after $K = \mathcal{O}\left(\frac{\sqrt{d}}{\epsilon}(\log(\frac{d}{\epsilon}))^2\right)$ iterations provided that $M \leq \frac{\epsilon}{\log(\sqrt{d}/\epsilon)}$ and $\eta \leq \frac{\epsilon}{\sqrt{d}\log(\sqrt{d}/\epsilon)}$.*

Proof. When $\beta(t) \equiv b$ for some $b > 0$, by Corollary D.4 and (77), we can compute that

$$
\begin{aligned}
\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0) \leq {} & \frac{\|\mathbf{x}_0\|_{L_2}}{\sqrt{m_0 e^{bK\eta} + 1 - m_0}} \\
& + \sum_{k=1}^{K} \frac{e^{(K-k)\eta\frac{1}{2}b + (K-k)\eta^2\max(1, L_0^2)b^2 + \frac{1}{2}L_1\eta^2(K-k)b}}{\left(m_0 e^{b(K-k)\eta} + 1 - m_0\right)^{\frac{1}{2}e^{-\frac{1}{2}\eta b}}} \\
& \cdot \left(\left(\frac{L_1}{2}\eta\left(1 + \|\mathbf{x}_0\|_{L_2} + \left(\|\mathbf{x}_0\|_{L_2}^2 + d\right)^{1/2}\right) + \frac{M}{2}\right) 2\left(e^{\frac{1}{2}b\eta} - 1\right)\right. \\
& + \sqrt{\eta}\max(1, L_0)b\left(e^{b\eta} - 1\right)^{1/2}\cdot\left(\left(\|\mathbf{x}_0\|_{L_2} + \left(\|\mathbf{x}_0\|_{L_2}^2 + d\right)^{1/2}\right)\left(\frac{1}{2} + \max(1, L_0)\right)b\eta\right. \\
& \left.\left. + (L_1 T + \|\nabla\log p_0(\mathbf{0})\|)\frac{1}{2}b\eta\right)\right),
\end{aligned}
$$

which implies that

$$
\begin{aligned}
\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0) \leq {} & \mathcal{O}\left(\frac{\sqrt{d}}{e^{\frac{1}{2}bK\eta}} + e^{K\eta^2\max(1, L_0^2)b^2 + \frac{1}{2}L_1\eta^2 Kb}\cdot e^{\frac{1}{2}bK\eta\left(e^{\frac{1}{2}\eta b} - 1\right)}\right. \\
& \left.\cdot K\left(\left(M + L_1\eta\sqrt{d}\right)\eta + \eta^2\left(\sqrt{d} + K\eta\right)\right)\right) \\
& \leq \mathcal{O}(\epsilon),
\end{aligned}
$$

and the condition $\eta \leq \bar{\eta}$ (where $\bar{\eta}$ is defined in (27)) holds provided that

$$K\eta = \frac{2}{b}\log\left(\frac{\sqrt{d}}{\epsilon}\right), \qquad M \leq \eta\sqrt{d}, \qquad \eta \leq \frac{\epsilon}{\sqrt{d}(K\eta)},$$

which implies that

$$M \leq \frac{\epsilon}{\log(\sqrt{d}/\epsilon)}, \qquad \eta \leq \frac{\epsilon}{\sqrt{d}\log(\sqrt{d}/\epsilon)}$$

and $K \geq \mathcal{O}\left(\frac{\sqrt{d}}{\epsilon}(\log(\frac{d}{\epsilon}))^2\right)$. This completes the proof. $\qquad\square$

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

We 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). Then we can obtain the following result from Corollary D.4.

Corollary D.6 (Restatement of Corollary 3.11)**.** *Assume* $\beta(t) = (b+at)^\rho$. *Then, we have* $\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0) \leq \mathcal{O}(\epsilon)$ *after* $K = \mathcal{O}\left(\frac{\sqrt{d}}{\epsilon}(\log(\frac{d}{\epsilon}))^{\frac{\rho+2}{\rho+1}}\right)$ *iterations provided that* $M \leq \frac{\epsilon}{\log(\sqrt{d}/\epsilon)}$ *and* $\eta \leq \frac{\epsilon}{\sqrt{d}\log(\sqrt{d}/\epsilon)}$.

Proof. When $\beta(t) = (b+at)^\rho$, by Corollary D.4 and (77), we can compute that

$$
\begin{aligned}
&\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0)\\
&\leq \frac{\sqrt{2d/m_0} + \|\mathbf{x}_*\|}{\sqrt{m_0 e^{\frac{1}{a(\rho+1)}((b+aK\eta)^{\rho+1}-b^{\rho+1})} + 1 - m_0}}\\
&\quad + \sum_{k=1}^{K} \frac{e^{\frac{1+L_1\eta}{2a(\rho+1)}((b+a(K-k)\eta)^{\rho+1}-b^{\rho+1})+\eta\max(1,L_0^2)\frac{1}{a(2\rho+1)}((b+a(K-k)\eta)^{2\rho+1}-b^{2\rho+1})}}{\left(e^{\frac{1}{a(\rho+1)}((b+a(K-k)\eta)^{\rho+1}-b^{\rho+1})} + 1 - m_0\right)^{\frac{1}{2}}e^{-\frac{1}{2}\eta(b+aK\eta)^\rho}}\\
&\quad \cdot \left(\left(\left(\frac{M}{2} + \frac{L_1}{2}\eta\left(1 + 2\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right) + \sqrt{d}\right)\right)\right.\right.\\
&\quad\quad \cdot 2\left(e^{\frac{1}{2}\frac{\left((b+a(K-k+1)\eta)^{\rho+1}-(b+a(K-k)\eta)^{\rho+1}\right)}{a(\rho+1)}} - 1\right)\\
&\quad + \sqrt{\eta}\left(\frac{1}{2} + \max(1, L_0)\right) \cdot \left(e^{\frac{\left((b+a(K-k+1)\eta)^{\rho+1}-(b+a(K-k)\eta)^{\rho+1}\right)}{a(\rho+1)}} - 1\right)^{1/2}\\
&\quad \cdot \left(\left(2\left(\sqrt{2d/m_0} + \|\mathbf{x}_*\|\right) + \sqrt{d}\right)\left(\frac{1}{2} + \max(1, L_0)\right)\frac{((b+a(K-k+1)\eta)^{\rho+1} - (b+a(K-k)\eta)^{\rho+1})}{a(\rho+1)}\right.\\
&\quad\quad \left.\left.\left. + (L_1K\eta + \|\nabla\log p_0(\mathbf{0})\|)\frac{1}{2}\frac{((b+a(K-k+1)\eta)^{\rho+1} - (b+a(K-k)\eta)^{\rho+1})}{a(\rho+1)}\right)\right)\right).
\end{aligned}
$$

We can compute that

$$
\begin{aligned}
&\mathcal{W}_2(\mathcal{L}(\mathbf{u}_K), p_0)\\
&\leq \mathcal{O}\Bigg(\frac{\sqrt{d}}{e^{\frac{1}{2a(\rho+1)}(b+aK\eta)^{\rho+1}}}\\
&\quad +\sum_{k=1}^{K}\frac{e^{\frac{1+L_1\eta}{2a(\rho+1)}((b+a(K-k)\eta)^{\rho+1}-b^{\rho+1})+\eta\max(1,L_0^2)\frac{1}{a(2\rho+1)}(b+aK\eta)^{2\rho+1}}}{\left(m_0 e^{\frac{1}{a(\rho+1)}((b+a(K-k)\eta)^{\rho+1}-b^{\rho+1})}+1-m_0\right)^{\frac{1}{2}}e^{-\frac{1}{2}\eta(b+aK\eta)^{\rho}}}\\
&\qquad \cdot\left(\left(M+L_1\eta\left(1+2\left(\sqrt{2d/m_0}+\|\mathbf{x}_*\|\right)+\sqrt{d}\right)\right)\left(e^{\frac{1}{2}(b+a(K-k+1)\eta)^{\rho}\eta}-1\right)\right.\\
&\qquad \left.+\sqrt{\eta}\left(e^{(b+a(K-k+1)\eta)^{\rho}\eta}-1\right)^{1/2}\cdot\left(\left(\sqrt{d}+K\eta\right)((K-k+1)\eta)^{\rho}\eta\right)\right)\Bigg)\\
&\leq \mathcal{O}\Bigg(\frac{\sqrt{d}}{e^{\frac{1}{2a(\rho+1)}(b+aK\eta)^{\rho+1}}}+e^{\eta\max(1,L_0^2)\frac{1}{a(2\rho+1)}(b+aK\eta)^{2\rho+1}}\cdot e^{\frac{1}{2a(\rho+1)}(b+aK\eta)^{\rho+1}\left(e^{\frac{1}{2}\eta(b+aK\eta)^{\rho}}-1\right)}\\
&\qquad \cdot K\left(\left(M+L_1\eta\sqrt{d}\right)(K\eta)^{\rho}\eta+\eta^2(K\eta)^{2\rho}\left(\sqrt{d}+K\eta\right)\right)\Bigg)\Bigg)\\
&\leq \mathcal{O}(\epsilon),
\end{aligned}
$$

and the condition $\eta \leq \bar{\eta}$ (where $\bar{\eta}$ is defined in (27)) holds provided that

$$
K\eta = \frac{(2a(\rho+1))^{\frac{1}{\rho+1}}}{a}\left(\log\left(\sqrt{d}/\epsilon\right)\right)^{\frac{1}{\rho+1}}-\frac{b}{a}, \qquad M \leq \eta\sqrt{d}, \qquad \eta \leq \frac{\epsilon}{\sqrt{d}(K\eta)^{\rho+1}},
$$

which implies that

$$
M \leq \frac{\epsilon}{\log(\sqrt{d}/\epsilon)}, \qquad \eta \leq \frac{\epsilon}{\sqrt{d}\log(\sqrt{d}/\epsilon)}
$$

and $K \geq \mathcal{O}\left(\frac{\sqrt{d}}{\epsilon}(\log(\frac{d}{\epsilon}))^{\frac{\rho+2}{\rho+1}}\right)$. This completes the proof. $\qquad\square$