Wasserstein Bounds for generative diffusion models with Gaussian tail targets

Anonymous authors
Paper under double-blind review

Abstract

We present an estimate of the Wasserstein distance between the data distribution and the generation of score-based generative models. The sampling complexity with respect to dimension is $\mathcal{O}(\sqrt{d})$, with a logarithmic constant. In the analysis, we assume a Gaussian-type tail behavior of the data distribution and an ϵ-accurate approximation of the score. Such a Gaussian tail assumption is general, as it accommodates practical target distributions derived from early stopping techniques with bounded support.

The crux of the analysis lies in the global Lipschitz bound of the score, which is shown from the Gaussian tail assumption by a dimension-independent estimate of the heat kernel. Consequently, our complexity bound scales linearly (up to a logarithmic constant) with the square root of the trace of the covariance operator, which relates to the invariant distribution of the forward process.

1 Introduction

Diffusion models (DM) are among the state-of-the-art tools in the new GenAI era. As generative models, diffusion models first link the target distribution to a distribution that is easy to sample via a diffusive process (forward). The generative (backward) process then *reverses* the diffusion, enabling samples from the easily sampled distribution to be transformed into samples from the target distribution. A well-known mathematical model that encapsulates this approach is the score-based stochastic differential equation (SDE) model (Song et al., 2020), where the forward and backward processes are represented by two SDEs that share the same marginal distribution (Anderson, 1982; Haussmann & Pardoux, 1986). In most cases, the forward process is assumed to be an Ornstein-Uhlenbeck (OU) process. The backward process incorporates the gradient of the logarithmic density (score) of the forward process. When the explicit form of the score is unknown, it is estimated by a neural network from discrete samples of the target distribution.

A major direction for the theoretical study of DMs is the convergence of the approximated backward process with limited data assumptions. Progress in this direction has been made, for example, in works such as (De Bortoli, 2023; Chen et al., 2023; Benton et al., 2023; Conforti et al., 2024; Gao & Zhu, 2024; Li & Yan, 2025; Silveri & Ocello, 2025). Essentially, there are two camps regarding the availability of the Lipschitz bound on the score function. When the Lipschitz bound of the score is available (regular target), the backward process is generally well-defined until $t = 0$, and convergence results are related to the Lipschitz bound, see e.g., (Chen et al., 2023; Conforti et al., 2024). Otherwise, when the bound is unavailable (singular target), the early stopping technique is introduced, and convergence results are related to the stopping time (De Bortoli, 2023; Lyu et al., 2022). Various analytical approaches have been adapted for these two types of assumptions.

In this work, we aim to present a general error analysis that applies to both regular and singular target distributions and also generalizes to an infinite-dimensional setting. The analysis is based on *a dimensionless and global-in-space Lipschitz bound* of the score, derived from a heat kernel estimation approach. Our complexity bounds, derived under the *Wasserstein-2-metric*, are *linear (with a logarithmic constant) in the square root of variance* of the Brownian motion in the forward process (Theorem 3.3). In the finite-dimensional case with the standard Gaussian as the base distribution (the invariant measure of the forward process), the variance is linear in the dimension and hence the complexity is $\mathcal{O}(\sqrt{d})$. For the infinite-dimensional case,

Table 1: Summary of previous bounds and our results for score-based diffusion models in d dimensions. The complexity bound gives the number of steps N to ensure error $\leq \epsilon_0$.

Target $\overrightarrow{P}_0$	Metric	Complexity	Result
Supp $B_R(0)$	$\mathcal{W}_1(\overrightarrow{P}_0, \overleftarrow{Q}_{T-\delta})$	$\log N = \mathcal{O}\Big(\frac{R^2(d+R^4)^2(\log R)^2}{\epsilon_0^2}\Big)$	(De Bortoli, 2023) Thm. 1 + Cor. 2
Supp $B_R(0)$	$\mathcal{W}_2(\overrightarrow{P}_0, \overleftarrow{Q}_{T-\delta})$	$\log N = \mathcal{O}\Big(\frac{R^2 d^2}{\epsilon_0^4}\Big)$	This work: Cor. 3.11
$\mathbb{E}\lvert X_0\rvert^2 < \infty$	$\mathrm{KL}(\overrightarrow{P}_\delta \lvert\lvert \overleftarrow{Q}_{T-\delta})$	$N = \mathcal{O}\Big(\frac{d \log^2 \frac{1}{\delta}}{\epsilon_0}\Big)$	(Benton et al., 2023) Cor. 1
$\mathbb{E}\lvert X_0\rvert^2 < \infty$	$\mathrm{TV}(\overrightarrow{P}_0 \lvert\lvert \overleftarrow{Q}_T)$	$N = \mathcal{O}\Big(\frac{d}{\epsilon_0}\Big)$	(Li & Yan, 2025) Thm. 1
$\nabla \log p_0$ L-lip	$\mathrm{KL}(\overrightarrow{P}_0 \lvert\lvert \overleftarrow{Q}_T)$	$N = \mathcal{O}\Big(\frac{d^2}{\epsilon_0}\Big)$	(Chen et al., 2023) Thm. 5
$\mathcal{F}(\overrightarrow{P}_0 \lvert \mathcal{N}(0, I_d)) \lesssim d$ $^{(*)}$	$\mathrm{KL}(\overrightarrow{P}_0 \lvert\lvert \overleftarrow{Q}_T)$	$N = \mathcal{O}\Big(\frac{d}{\epsilon_0^2} \log \frac{d}{\epsilon_0}\Big)$	(Conforti et al., 2024) Thm. 1
P_0 log-concave	$\mathcal{W}_2(\overrightarrow{P}_0, \overleftarrow{Q}_T)$	$N = \mathcal{O}\Big(\frac{d}{\epsilon_0^2} \log \frac{d}{\epsilon_0}\Big)$	(Gao et al., 2023) Tab. 2
P_0 one-side Lipschitz + weakly log-concave	$\mathcal{W}_2(\overrightarrow{P}_0, \overleftarrow{Q}_T)$	$N = \mathcal{O}\Big(\frac{d^2}{\epsilon_0^2}\Big)$	(Silveri & Ocello, 2025) Thm 3.5
G tail, Ass. 2	$\mathcal{W}_2(\overrightarrow{P}_0, \overleftarrow{Q}_T)$	$N = \mathcal{O}\Big(\frac{\sqrt{d}}{\epsilon_0} \log \frac{d}{\epsilon_0^2}\Big)$	This work: Cor. 3.4

* $\mathcal{F}$ denotes relative Fisher information. Our Gaussian tail assumption (Assumption 2) implies $\mathcal{F}(\overrightarrow{P}_0 \lvert \mathcal{N}(0, I_d)) \lesssim d$, and they are equivalent under the standard Gaussian case.

a Gaussian random field is taken as the base distribution, and the variance is linear with the trace of the covariance operator. It is worth noting that the results of this work adopt the *Wasserstein-2 distance* as the metric of error (instead of Kullback–Leibler (KL) divergence) due to its flexibility. More precisely, the reasons are two-fold: **(1)** In practical applications of diffusion models for the generation of structured data (image, text, video, protein, etc.), the target distributions mostly find their support in a compact sub-manifold, see further discussion of the manifold hypothesis in (Tenenbaum et al., 2000; Bengio et al., 2013; 2017). Consequently, under this hypothesis, the standard KL divergence cannot be consistently defined between the distribution obtained via the backward process, whose support is the entire space, and the target distribution with compact support. **(2)** In high (towards infinite) dimension settings, it becomes necessary to compactify the forward process. One way to achieve this is by choosing the covariance matrix (operator) C in the forward process (1) to have finite trace, so that the invariant distribution of the process has a finite second moment. The Wasserstein distance then scales with $\mathrm{Tr}(C)$, making it consistent with this compactification in the context of infinite-dimensional generative models. By contrast, the KL divergence scales with the ambient dimension and therefore cannot be directly extended to yield a dimensionless result. A motivating example is provided in Appendix B.1.

Related work

Complexity bounds (De Bortoli, 2022) established the first convergence guarantees in the 1-Wasserstein distance, assuming that the data distribution satisfies the so-called manifold hypothesis. Some recent works (Pierret & Galerne, 2024; Gao & Zhu, 2024; Gao et al., 2023; Silveri & Ocello, 2025; Li & Yan, 2025; Bruno & Sabanis, 2025; Beyler & Bach, 2025) established convergence guarantees under Gaussian, log-concave, weakly log-concave, or other minimal assumptions. We are also aware of several complexity bounds under KL divergence, for instance (Chen et al., 2023; Benton et al., 2023; Conforti et al., 2024). A common point of these approaches is the use of the chain rule for KL divergence, which follows from the Girsanov theorem, to separate the global error into local truncation errors. To bound local ones, the probabilistic viewpoint, which estimates the score function under the distribution of the forward process, comes into play. In this work, we instead provide a point-wise estimate of the score. Such an estimate also facilitates the analysis of the score under the approximated backward process, thereby providing a Wasserstein bound. In Table 1, we present the comparison of the complexity bounds.

Diffusion models in infinite dimension While diffusion models are usually defined in finite-dimensional spaces, many applications involve infinite-dimensional signals. In such cases, the observed data often consists of discrete samples from an underlying function. For instance, Bayesian inverse problems (Stuart, 2010). Up to now, there are many studies applying diffusion models to functional spaces (Lim et al., 2023; Kerrigan et al., 2023; Pidstrigach et al., 2023b; Franzese et al., 2024). While many theoretical studies (De Bortoli,

2023; Chen et al., 2023; Benton et al., 2023) suggest that performance guarantees deteriorate with increasing dimension. (Pidstrigach et al., 2023b) established a bound on the Wasserstein-2 Distance from the samples to the target distribution, which is dimension-independent but grows exponentially with the running time T. In this work, we further provide a Wasserstein-2-bound that is both dimension-independent and running time uniform under the Gaussian tail assumption.

Lipschitz bounds of the score The score functions in the SGMs are related to the gradient of log-density ($\log p$) of the forward process. It is well known that the function $\log p$ itself follows a viscous Hamilton-Jacobi (vHJ) equation, as seen in (8) in the later discussion. Then the Lipschitz bounds of the scores are equivalent to Hessian bounds for a vHJ equation. There are various regularity results in the literature for the original Fokker-Planck equation or the transformed vHJ, see (Fujita et al., 2006; Strömberg, 2010) and recent results in (Blessing & Kupper, 2022; Mooney et al., 2024). We would point out that, except (Mooney et al., 2024), most results are seeking a spatially global Hessian bound which only lasts for a finite time without the Gaussian tail assumption (Assumption 2) in this work. (Mooney et al., 2024) also provides a local-in-space and global-in-time bound, while only polynomial in dimension (d^3) complexity can be derived from it due to the spatial locality.

We are also aware of the literature on the topic of contraction properties or Lipschitz estimates of transport maps between measures.[1] The *Caffarelli's contraction theorem* (Caffarelli, 2000) in the optimal transport (OT) setup is the starting point, and see (Colombo et al., 2015) for generalization to Lipschitz estimate. Besides OT, the contractive map can also be attained by (reverse) heat flow (Kim & Milman, 2012). Recent generalizations to the Lipschitz estimates of the maps, including (Mikulincer & Shenfeld, 2023; Neeman, 2022; Fathi et al., 2024; Brigati & Pedrotti, 2024), obtain results for the boundedness of flow in a similar fashion to this work under diverse assumptions. In comparison, this work focuses more on deriving theories with applications to score-based diffusion models. In this regard, our assumptions cover the spatially anisotropic noise (no necessity for equivalence among A, C, and I_d in Assumption 2) and we discuss the convergence and complexity bounds in the discrete approximation of the generative flows.

The main contributions of this paper are:

- We introduce a Gaussian tail assumption (Assumption 2) and apply a heat kernel estimate (Theorem 3.1) inspired by the solution of the viscous Hamilton–Jacobi (vHJ) equation to obtain a spatially global Lipschitz constant for the modified score function that decays exponentially over time (Corollary 3.2). This assumption accommodates non-log-concave target distributions and covers practical scenarios, such as when the support of the target distribution is bounded and the early stopping technique (Lyu et al., 2022) is employed (Theorem 3.7).
- We establish a Wasserstein-2 bound (Theorem3.3) that depends only on the second moment of the base distribution, allowing the result to extend naturally to infinite-dimensional settings. When the second moment scales proportionally with the ambient dimension, the resulting iteration complexity achieves the state-of-the-art dependence on the dimension, with a rate of $\mathcal{O}(\sqrt{d})$. (Gao et al., 2023) Proposition 8 also shows that under the standard Gaussian distribution, such a complexity bound is optimal.
- As part of our methodological contribution, we provide a modified approach rather than conventional Lyapunov-type analysis, which utilizes the exponential decay structure of the score to show the uniform boundedness of accumulation error for an arbitrary long forward/backward process.

2 Preliminaries

This section contains the necessary notation and background to set the stage for the rest of the paper. We begin by introducing basic notations used throughout the text, followed by a review of the forward and backward processes in continuous time. We then discuss the discrete-time approximation commonly employed in diffusion models, together with its practical relevance in training score-based models. Finally, we return to

[1] More precisely, in the context of this work, the transport map refers to the push-forward map from the Gaussian (base distribution) to the target distribution p_0.

the continuous-time setting to present the viscous Hamilton–Jacobi (vHJ) formulation that lies at the heart of our regularity analysis.

General Notations Let γ_C be the Gaussian measure $\mathcal{N}(0, C)$. For an $n \times n$ matrix A, we use the operator norm $\|\cdot\|$,

$$\|A\| = \sup_{v \neq 0} \frac{|Av|}{|v|} := \text{the largest eigenvalue of } \sqrt{A^T A}.$$

For a symmetric, positive-definite $n \times n$ matrix A, we use $|\cdot|_A$ to denote weighted l_2 norm in $\mathbb{R}^n$ such that,

$$|x|_A^2 := \langle A^{-1/2}x, A^{-1/2}x \rangle.$$

When A is the identity matrix, we neglect the letter for simplicity and $|\cdot|$ is the standard l_2 norm in $\mathbb{R}^n$. For vector (matrix, correspondingly) valued function f with x as variable, $|f|_\infty := \sup_x |f(x)|$ (correspondingly, $\|f\|_\infty := \sup_x \|f(x)\|$).

2.1 Continuous-time formulation of diffusion models

A large class of generative diffusion models can be analyzed under the SDE framework (Song et al., 2020), which contains two processes: forward and backward. The forward process, which gradually transforms the data distribution into white noise, is an OU process as follows,

$$d\overrightarrow{X}_t = -\frac{1}{2}\overrightarrow{X}_t dt + \sqrt{C} dB_t, \quad 0 \leq t \leq T. \tag{1}$$

where B_t is a standard Brownian motion, C is a symmetric, positive-definite covariance matrix and T is the final time such that the distribution of X_T is close to the Gaussian distribution $\mathcal{N}(0, C)$, denoted as *base distribution*. The initial $\overrightarrow{X}_0$ follows the target(data) distribution, denoted as $\overrightarrow{P}_0$. We remark that in the majority of theories and applications of the diffusion model, C is assumed to be the identity. We maintain the spatially anisotropic noise assumption ($C \not\equiv I_d$) in the following derivation to enable our theories to generalize to an infinite-dimensional setting where some compactification is necessary, see also (Pidstrigach et al., 2023b). For compactness of the process, we made the following assumption to bound the second moment of the forward process.

Assumption 1. *The data distribution has a bounded second moment, $M_2 := \mathbb{E}_{\overrightarrow{P}_0} |x|^2 < \infty$. The covariance C in* (1) *is in trace-class. Furthermore, we denote,*

$$M_0 = \max\{\mathrm{Tr}(C), M_2, 1\}. \tag{2}$$

We denote the probability density of the forward process $\overrightarrow{X}_t$ by p_t, then p_t solves the Fokker-Planck equation with Cauchy data p_0:

$$\begin{cases} \partial_t p = \frac{1}{2}\big(\nabla \cdot (xp) + \nabla \cdot C\nabla p\big), \\ p(0, x) = p_0(x). \end{cases} \tag{3}$$

With time reversal $\overleftarrow{X}_t := \overrightarrow{X}_{T-t}$, the backward process $(\overleftarrow{X}_t)_{0 \leq t \leq T}$ satisfies the following SDE (Haussmann & Pardoux, 1986),

$$d\overleftarrow{X}_t = \big(\frac{1}{2}\overleftarrow{X}_t + s(T-t, \overleftarrow{X}_t)\big)dt + \sqrt{C} d\tilde{B}_t, \tag{4}$$

where $\tilde{B}_t$ is also a standard Brownian motion (may not be the same as B_t) and the term $s(t, x) := C\nabla \log p_t(x)$ is generally referred to as the score function. The process $(\overleftarrow{X}_t)_{0 \leq t \leq T}$ transforms noise into samples follows $\overrightarrow{P}_0$. We denote $\overrightarrow{P}_t$(correspondingly $\overleftarrow{P}_t$) as the marginal distribution of $\overrightarrow{X}_t$ in (1)($\overleftarrow{X}_t$ in (4)). Then, $\forall 0 \leq t \leq T$, $\overleftarrow{P}_t = \overrightarrow{P}_{T-t}$, especially $\overleftarrow{P}_T = \overrightarrow{P}_0$.

2.2 Training and discrete-time approximation of diffusion models

Since the closed form expression of p_0 is unknown, the score function $s(t,x) = C\nabla \log p_t(x)$ is not available. Thus, we model the score function by a neural network $s_\theta(t,x)$, where θ denotes latent variables of the neural network. We train the network by optimizing an L_2 estimation loss,

$$\mathbb{E}_{\overrightarrow{P}_t} \|s_\theta(t,x) - C\nabla \log p_t(x)\|^2.$$

Given the estimated score s_θ (assumed to be ϵ-accurate, specified in Assumption 3), one can generate samples of the target distribution p_0 by a numerical approximation of the backward process starting from the Gaussian distribution $\mathcal{N}(0,C)$,

$$d\overleftarrow{Y}_t = \big(\frac{1}{2}\overleftarrow{Y}_t + s_\theta(T-t, \overleftarrow{Y}_t)\big)dt + \sqrt{C}d\bar{B}_t. \tag{5}$$

In practice, diffusion models are trained and sampled in discrete time. Here, we introduce an Euler-type discretization of the continuous-time stochastic process, which facilitates the convergence analysis. Let $0 = t_0 \le t_1 \le \cdots \le t_N = T - \delta$ be the time discretization points (schedule), $\delta = 0$ for the normal setting and $\delta > 0$ for the early-stopping setting (Lyu et al., 2022), and we adopt the following discrete scheme. Starting from $\overleftarrow{Y}_0 \sim \mathcal{N}(0,C)$, for all $k = 0, 1, ..., N-1$,

$$\overleftarrow{Y}_{t_{k+1}} = \frac{1}{\sqrt{\alpha_k}}\left(\overleftarrow{Y}_{t_k} + (1-\alpha_k)s_\theta(T-t_k, \overleftarrow{Y}_{t_k})\right) + \sqrt{1-\alpha_k}\bar{z}_k, \tag{6}$$

where $\alpha_k = \exp(t_k - t_{k+1})$ and $\bar{z}_k \sim \mathcal{N}(0,C)$ are i.i.d.

Remark 2.1. The discretization (6) approximately corresponds to the discrete-time scheme introduced in (Ho et al., 2020); it can also be found in the analysis work (De Bortoli, 2023).

The bounds established later (Theorem 3.3) quantify how well the discrete process approximates the continuous-time backward SDE in Wasserstein distance. The step size $\tau := \max_k(t_{k+1} - t_k)$ enters the final error bound in the form $\mathrm{Tr}(C)\tau^2$, demonstrating the accumulation of numerical error in the generative process.

2.3 Foundational Ideas based on Heat Kernel Estimation

The convergence analysis of the discrete scheme (6) relies heavily on the regularity (up to second order) of the score potential function $\log p_t$, which follows a viscous Hamilton-Jacobi equation (vHJ). While similar derivations are known to experts (Evans, 2022), we provide a brief list of derivations here for completeness.

We first consider the following transform of $\log p_t$,

$$q(t,x) = -\log p_t(x) - \frac{x^T \bar{A}_t^{-1} x}{2}, \tag{7}$$

where $\bar{A}_t = Ae^{-t} + C(1-e^{-t})$. Then, q satisfies the vHJ,

$$\begin{cases} \partial_t q - \frac{1}{2}\nabla \cdot C\nabla q + \frac{1}{2}|\sqrt{C}\nabla q|^2 + (C\bar{A}_t^{-1} - \frac{1}{2}I_d)x \cdot \nabla q = \frac{1}{2}\mathrm{Tr}(C\bar{A}_t^{-1} - I_d), \\ q(0,x) = -h(x), \end{cases} \tag{8}$$

where $h(x) := \log p_0(x) + \frac{|x|_A^2}{2}$ is the non-Gaussian part of the log-likelihood function (the remainder term in later Assumption 2). To simplify (8), we let $f(t) = -\frac{1}{2}\int_0^t \mathrm{Tr}(C\bar{A}_s^{-1} - I_d)ds$ and make a two step change of variables: let $K(t) = (A\bar{A}_t^{-1})e^{-\frac{t}{2}}$, then $\bar{q}(t,x) = q(t, K(t)^{-1}\sqrt{C}x) + f(t)$ satisfies,

$$\begin{cases} \partial_t \bar{q} - \frac{1}{2}\nabla \cdot K(t)^2 \nabla \bar{q} + \frac{1}{2}|K(t)\nabla \bar{q}|^2 = 0, \\ \bar{q}(0,x) = \bar{h}(x) := -h(\sqrt{C}x). \end{cases} \tag{9}$$

Lastly, we define $\bar{p}(t,x) := e^{-\bar{q}(t,x)}$ that satisfies

$$\begin{cases} \partial_t \bar{p} - \frac{1}{2}\nabla \cdot K(t)^2 \nabla \bar{p} = 0, \quad \text{on } (0,\infty) \times \mathbb{R}^n, \\ \bar{p}(0,x) = e^{-\bar{h}(x)}. \end{cases} \tag{10}$$

(10) is a heat equation and admits the following solution from the heat kernel,

$$\bar{p}(t,x) = \frac{1}{(2\pi)^{\frac{n}{2}}} \int_{\mathbb{R}^n} \frac{1}{\sqrt{\det B(t)}} \exp\Big(\frac{-|x-y|^2_{B(t)}}{2}\Big) \exp\big(-\bar{h}(y)\big) dy, \tag{11}$$

where $B(t) := \int_0^t K(s)^2 ds = (e^{\frac{t}{2}} - e^{-\frac{t}{2}})K(t)$. It is worth mentioning that the kernel representation (11) leads directly to uniform-in-time bounds on $\nabla \bar{q}$ and $\nabla^2 \bar{q}$ (Theorem 3.1). These bounds yield the exponentially decaying Lipschitz constant of the modified score (Corollary 3.2), which is central to our Wasserstein convergence guarantees.

3 Results

In this section, we list the theoretical results, and their detailed proofs are provided in Appendix A. Section 3.1 is devoted to the Lipschitz bound of score with Gaussian tail assumption, which is based on a heat kernel estimation of (11). Section 3.2 lists the fundamental convergence result in the Wasserstein metric. Sections 3.3 and 3.4 present applications of our convergence result to the bounded-support assumption and Bayesian inverse problems, respectively.

3.1 Lipschitz Bound of Score Function

The foundation of our analysis is to provide a spatially global regularity estimate of p_t as the solution of an elliptic equation. These bounds control how sharply the true score can vary in space and directly impacts the stability and error propagation of the learned score in the generative process. Before obtaining such bounds, we need the following assumption to derive reasonable point-wise estimates of the gradients of the score function.

Assumption 2 (Gaussian tail)**.** *The density of target distribution $p_0 \in C^2(\mathbb{R}^d)$ and admits the tail decomposition,*

$$p_0(x) = \exp\Big(-\frac{|x|^2_A}{2}\Big) \exp\big(h(x)\big), \tag{12}$$

where

(i) A is a symmetric, positive-definite matrix which can be simultaneously diagonalized with C and satisfies $\|AC^{-1}\| < \infty$ and $\|CA^{-1}\| < \infty$,
(ii) The remainder term h satisfies $|\sqrt{C}\nabla h|_\infty < \infty$ and $\|C\nabla^2 h\|_\infty < \infty$.

Moreover, all the constants above are dimension-independent.

To the best of our knowledge, Assumption 2 in the form used here has not appeared explicitly in prior works on diffusion model theory, though it is closely related to conditions used in log-concavity and heat-flow regularity theory. Its advantage lies in yielding exponentially decaying Lipschitz/Hessian bounds for the modified score, which do not follow from classical Lipschitz or weak log-concavity assumptions; see the discussion below.

The Gaussian tail Assumption 2 is slightly more restrictive than both the L-Lipschitz condition on $\nabla \log p_0$ and the combination of weak log-concavity with one-sided Lipschitz-ness in (Silveri & Ocello, 2025) (Assumption H1). However, the L-Lipschitz condition on $\nabla \log p_0$ does not ensure a Lipschitz bound for $\nabla \log p_t$ (see Example 3.4 in (Mooney et al., 2024)), and Assumption H1 in (Silveri & Ocello, 2025) only guarantees an $\mathcal{O}(1)$ Hessian bound of $\log p_t$ over time, while the Gaussian tail yields an $\mathcal{O}(e^{-t})$ Hessian bound for the modified score function (Corollary 3.2) thus leads to improved complexity bounds (see Remark 3.5). Moreover, the Gaussian tail assumption does not require the target distribution to be log-concave, even when it is far from the origin.

To demonstrate the *practicality* of the Gaussian tail assumption, we provide two examples. First, in Theorem 3.7, we show that for any bounded-support target distribution, the density of the forward process at the early stopping time δ satisfies the assumption with $A = (1 - \exp(-\delta))I_d$. Second, the Gaussian

tail assumption also holds for certain posterior distributions arising in Bayesian inverse problems; see Theorem 3.12.

Under the Gaussian tail Assumption 2, we establish bounds on the Hessian and gradient of the transformed potential $\bar{q}(t,x)$ via the heat kernel estimation.

Theorem 3.1. *Under Assumption 2, the function $\bar{q}(t,x)$ in (9) satisfies, $\forall t \geq 0$,*

$$|\nabla \bar{q}(t,\cdot)|_\infty \leq |\nabla \bar{q}(0,\cdot)|_\infty, \quad \|\nabla^2 \bar{q}(t,\cdot)\|_\infty \leq \|\nabla^2 \bar{q}(0,\cdot)\|_\infty + |\nabla \bar{q}(0,\cdot)|_\infty^2. \tag{13}$$

Theorem 3.1 provides uniform in time estimates for the transformed variables in (9). By reversing the change of variables $(t,x) \to (t, (\bar{A}_t A^{-1}) e^{t/2} x)$ used to go from (8) to (9), these uniform bounds translate back to the original coordinates, resulting in the exponential decay estimates for the modified score function stated in the following corollary.

Corollary 3.2. *Define the modified score function as*

$$\tilde{s}(t,x) := C\nabla \log p_t(x) + C\bar{A}_t^{-1} x. \tag{14}$$

Suppose that Assumption 2 holds. Then, $\forall t \geq 0$, the following estimates hold:

$$\|\nabla \tilde{s}(t,\cdot)\|_\infty \leq e^{-t} L_0, \quad |\tilde{s}(t,\cdot)|_\infty \leq e^{-t/2} L_1, \tag{15}$$

where

$$\begin{cases} K = \max\{1, \|AC^{-1}\|\}, \\ L_0 = K^2\big(\|C\nabla^2 h\|_\infty + |\sqrt{C}\nabla h|_\infty^2\big), \\ L_1 = K\|C\|^{1/2}|\sqrt{C}\nabla h|_\infty. \end{cases}$$

Detailed proofs can be found in Appendices A.1 and A.2. To be noted, K, L_0, L_1 are bounded dimension-free constants. For later discussion, we also denote,

$$L_2 := \max\{\|I - CA^{-1}\|, \|AC^{-1} - I\|\}.$$

This result provides a formal justification for a well-known empirical feature: *the score function becomes smoother at higher noise levels.* This smoothing is crucial because (i) it enables stable training at early timesteps, and (ii) it ensures that discretization error does not explode when simulating the backward process.

3.2 Main Convergence theories

In the preceding section, we established the crucial analytical properties of the modified score function $\tilde{s}(t,x)$ that drives the reverse diffusion process, namely the *exponential decay of its Lipschitz constant.* This section utilizes these theoretical foundations to provide rigorous, quantifiable bounds on the convergence of our discrete sampling scheme (6) under the Wasserstein-2-distance,

$$\mathcal{W}_2(\mu,\nu) := \left(\inf_\pi \int |x-y|^2 d\pi(x,y)\right)^{1/2},$$

where π runs over all measures which have marginals μ and ν. Our objective is to clearly dissect the sources of the total sampling error and demonstrate how this error decomposes and evolves with respect to the diffusion time T, the step size τ, and the score network approximation accuracy ϵ. To achieve this, we first introduce the following assumptions concerning the learned score function s_θ. Assumption 3 quantifies how accurately the learned score approximates the true score at discretization points. Assumption 4 ensures a Lipschitz-type regularity for the learned score, consistent with the theoretically guaranteed regularity of the true score.

Assumption 3. *s_θ is an ϵ accurate approximation to s on average over the discretization points, i.e.,*

$$\frac{1}{T}\sum_{k=0}^{N-1}(t_{k+1}-t_k)\mathbb{E}\Big|s(T-t_k, \overrightarrow{X}_{T-t_k}) - s_\theta(T-t_k, \overrightarrow{X}_{T-t_k})\Big|^2 \leq \epsilon^2. \tag{16}$$

We would like to point out that, due to the limited access to the score approximation error, when deriving the complexity bounds, we always assume a sufficiently accurate score approximation.

Assumption 4. *Denote $\tilde{s}_\theta(t,x) = s_\theta(t,x) + C\bar{A}_t^{-1}x$, we assume that $\tilde{s}_\theta(t,x)$ shares the same properties as $\tilde{s}(t,x)$ in Corollary 3.2 at the time discretization points, i.e.*

$$\forall k \in \{0,1,\dots,N-1\}, \quad \|\nabla \tilde{s}_\theta(T-t_k\cdot)\|_\infty \le L_0 e^{-T+t_k}. \tag{17}$$

The Assumption 4 can be relaxed to require that the total discrete-time sum of the score gradient is bounded, see Remark 3.6.

Now we present the main theorem in this part.

Theorem 3.3. *Suppose Assumptions 1, 2, 3, and 4 hold, with the step size $\tau := \sup_k\{t_{k+1}-t_k\}$ sufficiently small. Then sampling via scheme* (6) *yields*

$$\mathcal{W}_2^2(\overrightarrow{P}_0, \overleftarrow{Q}_T) \le K_1\Big(e^{-2T}\left(M_2 + \mathrm{Tr}(C)\right) + \epsilon^2 T + \mathrm{Tr}(C)\tau^2\Big), \tag{18}$$

where $K_1 = e^{4+3(L_0+L_2)}\max\{(e+1)e, e^3(L_0+L_2)^2\}$ is a dimension-free constant.

Proof see Appendix A.3. This theorem quantifies how the sampling error decomposes into: (i) a term e^{-2T} from incomplete denoising, (ii) a term $\epsilon^2 T$ due to score approximation error, (iii) a discretization error $\mathrm{Tr}(C)\tau^2$. Each term has a direct and clear interpretation for practitioners choosing T, network size, and timestep schedules. As a direct consequence of Theorem 3.3,

Corollary 3.4. *Under the same assumptions as Theorem 3.3, the following complexity bound guarantees that the distribution $\overleftarrow{Q}_T$ satisfies $\mathcal{W}_2(\overrightarrow{P}_0, \overleftarrow{Q}_T) = \mathcal{O}(\epsilon_0)$:*

$$T = \mathcal{O}\left(\log\frac{M_2+\mathrm{Tr}(C)}{\epsilon_0^2}\right), \quad N = \mathcal{O}(\frac{T}{\tau}) = \mathcal{O}\left(\frac{T}{\epsilon_0}\sqrt{\mathrm{Tr}(C)}\right). \tag{19}$$

Fixing the second moment M_2 and trace of covariance matrix C of base distribution, the complexity bound in Corollary 3.4 does not depend on the dimension and hence can be generalized to some infinite-dimensional generative models; see Theorem B.3 and related discussion in Appendix B.

We now provide two remarks: the first concerns the optimality of the result, and the latter addresses the relaxation of assumptions on score approximations.

Remark 3.5. When assuming the second moment of target distribution M_2 and the trace of the diffusion covariance $\mathrm{Tr}(C)$ scale linearly with space dimension d, the complexity bound in Corollary 3.4 is $\mathcal{O}(\sqrt{d})$ with a logarithmic constant. In (Gao et al., 2023), Proposition 8 shows that, with the standard Gaussian as the target distribution, such a complexity bound is optimal. Notably, in the line of pursuing complexity bounds under more general assumptions (than log-concaveness), a very recent work (Silveri & Ocello, 2025) obtained an $\mathcal{O}(d^2)$ bound[2] through the weakly log-concave profile propagation framework established in (Saremi et al., 2023; Conforti, 2024; Conforti et al., 2025). In Table 1, we provide a non-exhaustive list of complexity bounds for comparison.

Remark 3.6. Assumption 4 can be relaxed as follows. We only require that for all $k \in \{0,1,\dots,N-1\}$,

$$\sum_{k=0}^{N-1}(t_{k+1}-t_k)\|\nabla\tilde{s}_\theta(T-t_k,\cdot)\|_\infty \le B, \tag{20}$$

for some constant $B > 0$ independent of T. Under this condition, Theorem 3.3 still holds with the constant term K_1 in (18) replaced by,

$$K_1' = e^{4+3(L_2+eB)}\max\{(e+1)e, e^3(L_0+L_2)^2\}.$$

See Appendix A.4.

[2]In arriving at this, we also assume the second moment scales linearly with dimension in their Theorem 3.5.

3.3 Convergence for bounded-support target

This subsection specializes the convergence theory of Theorem 3.3 to an important and practically relevant setting: when the target distribution p_0 is supported on a bounded set. Such distribution widely applies to various practical applicative scenario of generative models and can be derived from the manifold hypothesis (Bengio et al., 2013). We make this assumption explicit as follows.

Assumption 5. *The target distribution $\overrightarrow{P}_0$ is supported on a bounded set, i.e., there exists $R < \infty$ such that*

$$\mathrm{Diam}(\mathrm{Supp}\,\overrightarrow{P}_0) \le R.$$

While Assumption 5 alone does not impose any smoothness or tail behavior on p_0, a key phenomenon is that the forward Ornstein–Uhlenbeck diffusion instantaneously regularizes such distributions. In particular, for every $t > 0$, the density p_t acquires Gaussian-type tails and thus satisfies the Gaussian tail condition of Assumption 2. To quantify this regularizing effect, we first record a general result on Gaussian convolution of a bounded-support distribution.

Theorem 3.7. *Let $Q_\sigma = \mathcal{N}(0, \sigma^2 I_d) * Q_0$, where Q_0 follows Assumption 5. Then Q_σ has smooth density q_σ and we define*

$$g(x) := \log q_\sigma(x) + \frac{|x|^2}{2\sigma^2}.$$

It follows that,

$$|\nabla g|_\infty \le \frac{R}{\sigma^2}, \quad \|\nabla^2 g\|_\infty \le \frac{2R^2}{\sigma^4}. \tag{21}$$

Proof see Appendix A.5.

Similar estimates in Theorem 3.7 can be found in (De Bortoli, 2022; Mooney et al., 2024). In the current form (21), we extract the spatial growing part to ensure uniform boundedness in space to be consistent with Assumption 2.

Applying Theorem 3.7 to the forward OU-process (1) with covariance $C = I_d$ and stopping time δ, we obtain that the marginal $\overrightarrow{P}_\delta$ has smooth density p_δ and satisfies,

$$p_\delta(x) = \exp\Big(-\frac{|x|^2}{2(1-e^{-\delta})}\Big)\exp\big(h(x)\big),$$

and

$$|\nabla h|_\infty \le \frac{e^{-\frac{\delta}{2}}R}{1-e^{-\delta}}, \qquad \|\nabla^2 h\|_\infty \le 2\frac{e^{-\delta}R^2}{(1-e^{-\delta})^2}.$$

Hence p_δ satisfies Assumption 2, and the score Lipschitz bounds from Corollary 3.2 apply directly.

Corollary 3.8. *Suppose that $\overrightarrow{P}_0$ follows the Assumption 5. Let $C = I_d$ and $A = (1-e^{-\delta})I_d$. Then, the corresponding $\bar{A}_t = (1-e^{-t})I_d$ and for all $t \ge \delta$, we have,*

$$\|\nabla^2 \log p_t(x) + \bar{A}_t^{-1}\|_\infty \le 3\Big(\frac{R}{1-e^{-\delta}}\Big)^2 e^{-t}, \quad |\nabla \log p_t(x) + \bar{A}_t^{-1}x|_\infty \le \frac{R}{1-e^{-\delta}}e^{-\frac{t}{2}}.$$

Proof see Appendix A.2. Using the Lipschitz bound in Corollary 3.8, we obtain a Wasserstein-2 distance bound analogous to Theorem 3.3.

Theorem 3.9. *Suppose that $\overrightarrow{P}_0$ follows the Assumption 5, the early stopping time $\delta \le 1$, and the Assumption 3 holds. Then, sampling via scheme (6) with sufficiently small step size τ yields,*

$$\mathcal{W}_2^2(\overrightarrow{P}_\delta, \overleftarrow{Q}_{T-\delta}) \le K_2(R,\delta)\left(e^{-2T+\delta}\left(R^2+d\right) + \epsilon^2 T + d\tau^2\right), \tag{22}$$

where $K_2(R,\delta) = 2\exp\left(7 + \frac{12R^2}{(1-e^{-\delta})^2} + \frac{4}{1-e^{-\delta}}\right)$.

Proof see Appendix A.6.

With the result in Theorem 3.9, we can get the complexity bound with early stopping setting under the bounded-support assumption by direct computation.

Corollary 3.10. *Under the same assumption as Theorem 3.9, achieving a distribution* $\overleftarrow{Q}_{T-\delta}$ *such that* $\mathcal{W}_2(\overrightarrow{P}_\delta, \overleftarrow{Q}_{T-\delta}) = \mathcal{O}(\epsilon_0)$ *requires:*

$$T = \mathcal{O}\Big(\frac{R^2}{\delta^2} + \log\frac{R^2+d}{\epsilon_0^2}\Big), \quad N = \frac{T-\delta}{\tau} = \mathcal{O}\left(\frac{\sqrt{d}T}{\epsilon_0}\exp\left(\frac{2}{1-e^{-\delta}} + \frac{6R^2}{(1-e^{-\delta})^2}\right)\right).$$

When fixing the early stopping time δ *and support radius* R*,*

$$N = \mathcal{O}\Big(\frac{\sqrt{d}}{\epsilon_0}\log\frac{R^2+d}{\epsilon_0^2}\Big). \tag{23}$$

Noticing that, $\mathcal{W}_2(\overrightarrow{P}_\delta, \overrightarrow{P}_0) \leq \sqrt{\mathbb{E}|\overrightarrow{X}_\delta - \overrightarrow{X}_0|^2} \leq 2\sqrt{M_0\delta}$, we have the following complexity bound with respect to $\overrightarrow{P}_0$.

Corollary 3.11. *Under the same assumption as Theorem 3.9, reaching a distribution* $\overleftarrow{Q}_{T-\delta}$ *such that* $\mathcal{W}_2(\overrightarrow{P}_0, \overleftarrow{Q}_{T-\delta}) = \mathcal{O}(\epsilon_0)$ *requires,*

$$\delta = \mathcal{O}\Big(\frac{\epsilon_0^2}{M_0}\Big), \quad \log N = \mathcal{O}\Big(\frac{R^2}{\delta^2}\Big) = \mathcal{O}\Big(\frac{R^2M_0^2}{\epsilon_0^4}\Big). \tag{24}$$

The logarithmic complexity $\log N = \mathcal{O}(\frac{R^2}{\delta^2})$ arises jointly from two factors: the $\frac{R^2}{(1-e^{-\delta})^2}$ dependence in the Hessian bound of the early-stopped log density (see Corollary 3.8), and the accumulation of error with respect to the exponential of the Lipschitz-ness of the flow, which is tied to this Hessian bound. To the best of our knowledge, such a dependence on δ is non-improvable without additional geometric assumptions on the support of $\overrightarrow{P}_0$, see (Mooney et al., 2024). Similar exponential dependencies also appear in (De Bortoli, 2023), while our dependence on R is improved; see Table 1.

3.4 Convergence in the Bayesian Inverse problems

Another potential application of generative models is to generate the posterior distribution in Bayesian inverse problems. See (Stuart, 2010) for a detailed review. Here, we restrict our theories to the following type of applicative scenario, where we consider a non-linear observation $G \in C_2^b(\mathbb{R}^d, \mathbb{R}^m)$ from the state space $\mathbb{R}^d$ and its observation $y \in \mathbb{R}^m$. We further assume that the prior of the state and the observation noise distribution follow Gaussian-type distributions with C and Σ denoting the covariance matrices, respectively. Then the posterior of the state, also the target distribution of the generative model, follows,

$$p_0(x) = D_0 \exp\Big(-\frac{|x|_C^2}{2}\Big)\exp\Big(-\frac{|G(x)-y|_\Sigma^2}{2}\Big), \tag{25}$$

where D_0 is some normalizing constant.

To construct a generative model to sample the posterior, we take the covariance matrix of the base distribution in (1) identical to the prior covariance matrix C in (25). A conditioned score (referred to as the conditional de-noising estimator in the literature (Batzolis et al., 2021)) is trained with the following loss,

$$\mathbb{E}_{p_t(x;y)}|s_\theta(t,x;y) - C\nabla_x \log p_t(x;y)|^2,$$

where p_t is then the joint distribution of (X_t, Y) in which Y follows $G(X_0) + \mathcal{N}(0,\Sigma)$. For the generation process of the posterior distribution with observation y, we assume the estimated score $s_\theta(t,x;y)$ satisfies Assumption 3 and 4. Then, we have the following theorem.

Theorem 3.12. *Suppose that the target distribution admits the density in* (25) *and Assumption 1 holds. Then, sampling via scheme* (6) *with sufficiently small step size* τ *yields,*

$$\mathcal{W}_2^2\big(\overrightarrow{P}_0(\cdot;y),\overleftarrow{Q}_T(\cdot;y)\big) \leq K_3\Big(e^{-2T}\big(M_2+\mathrm{Tr}(C)\big)+\epsilon^2 T+\mathrm{Tr}(C)\tau^2\Big), \tag{26}$$

where K_3 *is a dimensionless constant determined by* $(\|C\|,\|\Sigma\|,G,y)$*, the explicit expression is provided in Table 2.*

Proof see Appendix A.7.

This provides a theoretical grounding for the growing empirical use of diffusion models as posterior samplers (e.g. conditional diffusion models), justifying that with sufficiently accurate score learning, the generated posterior approximations are provably close to the true Bayesian posterior in the Wasserstein-2-distance.

4 Conclusion

In this work, we presented a unified and generalizable theoretical framework for analyzing the convergence of score-based generative diffusion models (SGMs), specifically focusing on bounding the Wasserstein-2 distance between the target and generated distributions. Our analysis hinges on the introduction of the *Gaussian-tail assumption* (Assumption 2), a more general condition that accommodates practical scenarios such as distributions with bounded support and those arising from early stopping techniques. A core technical contribution is establishing the *exponential decay over time* of the Lipschitz constant for the modified score function (Corollary 3.2), derived via a dimension-independent heat kernel estimate. This enabled us to prove a rigorous Wasserstein-2 distance bound (Theorem 3.3) that only depends on the second moment of the base distribution, thereby naturally extending the analysis to the infinite-dimensional setting. Furthermore, for the standard finite-dimensional case, our analysis demonstrates an optimal iterative complexity rate of $\mathcal{O}(\sqrt{d})$ (Corollary 3.4), matching the current state-of-the-art. Overall, our work provides stringent theoretical guarantees, affirming the convergence and optimal dimension-dependence of the discrete approximation scheme for SDE-based score models under a more general class of target distributions, judged by the Wasserstein-2 metric.

Moving forward, a compelling direction for future research involves investigating the higher-order regularity of the score function. While our current analysis provides regularity estimates up to the second derivative—corresponding to the second-order boundedness of the Gaussian-tail's remainder term—we believe that imposing higher-order regularity on this remainder term (which is feasible in practice via bounded support combined with early stopping) could yield analogous higher-order regularity estimates for the score function itself. Such advanced regularity understanding is beneficial for guiding neural network training objectives and enabling the discovery of higher-order numerical SDE schemes, thereby significantly improving the convergence rate and accuracy of the discrete numerical scheme relative to the true backward diffusion process.

Broader Impact

The paper provides insights into the regularity and convergence of one of the state-of-the-art generative models, which the GenAI community may benefit from. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

B. Anderson. Reverse-time diffusion equation models. *Stoch. Process. Appl.*, 12(3):313–326, 1982.

Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Yoshua Bengio, Ian Goodfellow, and Aaron Courville. *Deep learning*, volume 1. MIT press Cambridge, MA, USA, 2017.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*, 2023.

Eliot Beyler and Francis Bach. Convergence of deterministic and stochastic diffusion-model samplers: A simple analysis in wasserstein distance. *arXiv preprint arXiv:2508.03210*, 2025.

Jonas Blessing and Michael Kupper. Viscous hamilton–jacobi equations in exponential orlicz hearts. *Journal de Mathématiques Pures et Appliquées*, 163:654–672, 2022. ISSN 00217824. doi: 10.1016/j.matpur.2022.05.018. URL `https://linkinghub.elsevier.com/retrieve/pii/S002178242200071X`.

Giovanni Brigati and Francesco Pedrotti. Heat flow, log-concavity, and lipschitz transport maps. *arXiv preprint arXiv:2404.15205*, 2024.

Stefano Bruno and Sotirios Sabanis. Wasserstein convergence of score-based generative models under semiconvexity and discontinuous gradients. *arXiv preprint arXiv:2505.03432*, 2025.

L. Caffarelli. Monotonicity properties of optimal transportation¶ and the fkg and related inequalities. *Communications in Mathematical Physics*, 214:547–563, 2000.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions, 2023. URL `https://arxiv.org/abs/2211.01916`.

Maria Colombo, Alessio Figalli, and Yash Jhaveri. Lipschitz changes of variables between perturbations of log-concave measures. *arXiv preprint arXiv:1510.03687*, 2015.

Giovanni Conforti. Weak semiconvexity estimates for schrödinger potentials and logarithmic sobolev inequality for schrödinger bridges. *Probability Theory and Related Fields*, 189(3):1045–1071, 2024.

Giovanni Conforti, Alain Durmus, and Marta Gentiloni Silveri. Kl convergence guarantees for score diffusion models under minimal data assumptions, 2024. URL `https://arxiv.org/abs/2308.12240`.

Giovanni Conforti, Daniel Lacker, and Soumik Pal. Projected langevin dynamics and a gradient flow for entropic optimal transport. *Journal of the European Mathematical Society*, 2025.

Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic equations in infinite dimensions*, volume 152. Cambridge university press, 2014.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis, May 2023. URL `http://arxiv.org/abs/2208.05314`. arXiv:2208.05314 [cs, math, stat].

Lawrence C Evans. *Partial differential equations*, volume 19. American Mathematical Society, 2022.

Max Fathi, Dan Mikulincer, and Yair Shenfeld. Transportation onto log-lipschitz perturbations. *Calculus of Variations and Partial Differential Equations*, 63(3):61, 2024.

Giulio Franzese, Giulio Corallo, Simone Rossi, Markus Heinonen, Maurizio Filippone, and Pietro Michiardi. Continuous-time functional diffusion processes. *Advances in Neural Information Processing Systems*, 36, 2024.

Yasuhiro Fujita, Hitoshi Ishii, and Paola Loreti. Asymptotic solutions of viscous hamilton–jacobi equations with ornstein–uhlenbeck operator. *Communications in Partial Differential Equations*, 31(6):827–848, 2006.

Xuefeng Gao and Lingjiong Zhu. Convergence analysis for general probability flow odes of diffusion models in wasserstein distances, 2024. URL `https://arxiv.org/abs/2401.17958`.

Xuefeng Gao, Hoang M. Nguyen, and Lingjiong Zhu. Wasserstein convergence guarantees for a general class of score-based generative models, 2023. URL `https://arxiv.org/abs/2311.11003`.

Ulrich G Haussmann and Etienne Pardoux. Time reversal of diffusions. *The Annals of Probability*, pp. 1188–1205, 1986.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Ioannis Karatzas and Steven Shreve. *Brownian motion and stochastic calculus*, volume 113. springer, 2014.

Gavin Kerrigan, Justin Ley, and Padhraic Smyth. Diffusion Generative Models in Infinite Dimensions, February 2023. URL `http://arxiv.org/abs/2212.00886`. arXiv:2212.00886 [cs, stat].

Young-Heon Kim and Emanuel Milman. A generalization of caffarelli's contraction theorem via (reverse) heat flow. *Mathematische Annalen*, 354(3):827–862, 2012. ISSN 0025-5831, 1432-1807. doi: 10.1007/s00208-011-0749-x. URL `http://link.springer.com/10.1007/s00208-011-0749-x`.

Gen Li and Yuling Yan. O(d/t) convergence theory for diffusion probabilistic models under minimal assumptions, 2025. URL `https://arxiv.org/abs/2409.18959`.

Jae Hyun Lim, Nikola B. Kovachki, Ricardo Baptista, Christopher Beckham, Kamyar Azizzadenesheli, Jean Kossaifi, Vikram Voleti, Jiaming Song, Karsten Kreis, Jan Kautz, Christopher Pal, Arash Vahdat, and Anima Anandkumar. Score-based Diffusion Models in Function Space, November 2023. URL `http://arxiv.org/abs/2302.07400`. arXiv:2302.07400 [cs, math, stat].

Zhaoyang Lyu, Xudong Xu, Ceyuan Yang, Dahua Lin, and Bo Dai. Accelerating diffusion models via early stop of the diffusion process. *arXiv preprint arXiv:2205.12524*, 2022.

Dan Mikulincer and Yair Shenfeld. On the lipschitz properties of transportation along heat flows. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2020-2022*, pp. 269–290. Springer, 2023.

Connor Mooney, Zhongjian Wang, Jack Xin, and Yifeng Yu. Global well-posedness and convergence analysis of score-based generative models via sharp lipschitz estimates. *arXiv preprint arXiv:2405.16104*, 2024.

Joe Neeman. Lipschitz changes of variables via heat flow. *arXiv preprint arXiv:2201.03403*, 2022.

Jakiw Pidstrigach, Youssef Marzouk, Sebastian Reich, and Sven Wang. Infinite-dimensional diffusion models. *arXiv preprint arXiv:2302.10130*, 2023a.

Jakiw Pidstrigach, Youssef Marzouk, Sebastian Reich, and Sven Wang. Infinite-Dimensional Diffusion Models, October 2023b. URL `http://arxiv.org/abs/2302.10130`. arXiv:2302.10130 [cs, math, stat].

Emile Pierret and Bruno Galerne. Diffusion models for gaussian distributions: Exact solutions and wasserstein errors, 2024. URL `https://arxiv.org/abs/2405.14250`.

Saeed Saremi, Ji Won Park, and Francis Bach. Chain of log-concave markov chains. *arXiv preprint arXiv:2305.19473*, 2023.

Marta Gentiloni Silveri and Antonio Ocello. Beyond log-concavity and score regularity: Improved convergence bounds for score-based generative models in w2-distance. In *Forty-second International Conference on Machine Learning*, 2025.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Thomas Strömberg. Semiconcavity estimates for viscous hamilton–jacobi equations. *Archiv der Mathematik*, 94(6), 2010.

Andrew M Stuart. Inverse problems: a bayesian perspective. *Acta numerica*, 19:451–559, 2010.

Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Appendix

The appendix consists of three parts. In Section A, we present all the detailed proofs. In Section B, we discuss generalizing our theories to infinite dimensions. In Section C, we provide the explicit expressions of the constants appearing in the context.

A Proofs of Theorems

Here we present detailed proofs.

A.1 Proof of Theorem 3.1 (heat kernel estimation)

Proof. Consider the solution of (10) given by (11),

$$\bar{p}(t,x) = \frac{1}{(2\pi)^n}\int_{\mathbb{R}^n}\frac{1}{\sqrt{\det B(t)}}\exp\Big(\frac{-|x-y|^2_{B(t)}}{2}\Big)\exp\big(-\bar{h}(y)\big)dy, \quad (t,x)\in(0,\infty)\times\mathbb{R}^n.$$

So $\bar{q}(t,x) = -\log\bar{p}(t,x)$ satisfies,

$$\begin{aligned}\nabla_x\bar{q}(t,x) &= -\frac{\nabla_x\bar{p}(t,x)}{\bar{p}(t,x)} = -\frac{\int_{\mathbb{R}^n}\big(\nabla_x\exp(\frac{-|x-y|^2_{B(t)}}{2})\big)\exp\big(-\bar{h}(y)\big)dy}{\int_{\mathbb{R}^n}\exp(\frac{-|x-y|^2_{B(t)}}{2})\exp\big(-\bar{h}(y)\big)dy}\\ &= \frac{\int_{\mathbb{R}^n}\big(\nabla_y\exp(\frac{-|x-y|^2_{B(t)}}{2})\big)\exp\big(-\bar{h}(y)\big)dy}{\int_{\mathbb{R}^n}\exp(\frac{-|x-y|^2_{B(t)}}{2})\exp\big(-\bar{h}(y)\big)dy}\\ &= \frac{\int_{\mathbb{R}^n}\big(\nabla_y\bar{h}(y)\big)\exp(\frac{-|x-y|^2_{B(t)}}{2})\exp\big(-\bar{h}(y)\big)dy}{\int_{\mathbb{R}^n}\exp(\frac{-|x-y|^2_{B(t)}}{2})\exp\big(-\bar{h}(y)\big)dy}.\end{aligned}$$

Here, the third line is derived from the integration by parts formula.

Since $\exp(\frac{-|x-y|^2_{B(t)}}{2})\exp\big(-\bar{h}(y)\big)\geq 0$, taking absolute value we get $|\nabla_x\bar{q}(t,x)|\leq|\nabla\bar{h}|_\infty$, thus,

$$|\nabla\bar{q}(t,\cdot)|_\infty \leq |\nabla\bar{h}|_\infty = |\nabla\bar{q}(0,\cdot)|_\infty.$$

For any unit direction z, we denote ∇_z as taking derivative along that direction and ∇_z as taking derivative twice along that direction, i.e., for a given C^2 function f,

$$\nabla_z f(x) = \langle z, \nabla f(x)\rangle, \quad \nabla_z^2 f(x) = \langle z, \nabla^2 f(x) z\rangle,$$

where $\nabla f(x)$ is the gradient of f and $\nabla^2 f(x)$ is the Hessian matrix of f at x. Using the same method as above, we get,

$$\begin{aligned}\nabla_z^2\bar{q}(t,x) = &\frac{\int_{\mathbb{R}^n}\big(\nabla_z^2\bar{h}-(\nabla_z\bar{h})^2\big)(y)\exp(\frac{-|x-y|^2_{B(t)}}{2})\exp\big(-\bar{h}(y)\big)dy}{\int_{\mathbb{R}^n}\exp(\frac{-|x-y|^2_{B(t)}}{2})\exp\big(-\bar{h}(y)\big)dy}\\ &+\Big(\frac{\int_{\mathbb{R}^n}\big(\nabla_z\bar{h}(y)\big)\exp(\frac{-|x-y|^2_{B(t)}}{2})\exp\big(-\bar{h}(y)\big)dy}{\int_{\mathbb{R}^n}\exp(\frac{-|x-y|^2_{B(t)}}{2})\exp\big(-\bar{h}(y)\big)dy}\Big)^2.\end{aligned}$$

Taking absolute value again, we get,

$$|\nabla_z^2\bar{q}(t,x)| \leq |\nabla_z^2\bar{h}|_\infty + |\nabla_z\bar{h}|^2_\infty \leq \|\nabla^2\bar{h}\|_\infty + |\nabla\bar{h}|^2_\infty, \quad \forall x\in\mathbb{R}^n,$$

thus,

$$\|\nabla^2\bar{q}(t,x)\| = \sup_{|z|=1}|\nabla_z^2\bar{q}(t,x)| \leq |\nabla^2\bar{h}\|_\infty + |\nabla\bar{h}|^2_\infty, \quad \forall x\in R^n.$$

Since the bound for $\|\nabla^2\bar{q}(t,x)\|$ does not rely on x, we get,

$$\|\nabla^2\bar{q}(t,\cdot)\|_\infty \le \|\nabla^2\bar{h}\|_\infty + |\nabla\bar{h}|_\infty^2 = \|\nabla^2\bar{q}(0,\cdot)\|_\infty + |\nabla\bar{q}(0,\cdot)|_\infty^2. \tag{27}$$

□

The above analysis is developed to facilitate the case $C \neq I_d$; other standard PDE approaches, like the classical Bernstein method, can also be applied for the isotropic case, i.e., $C = I_d$.

A.2 Proof of Corollary 3.2 and Corollary 3.8

Proof. Recall that $\bar{q}(t,x) = q(t, K(t)^{-1}\sqrt{C}x) + f(t)$, so,

$$\begin{aligned} K(t)\sqrt{C^{-1}}\nabla\bar{q}(t,x) =& \nabla q(t, K(t)^{-1}\sqrt{C}x), \\ K(t)^2 C^{-1}\nabla^2\bar{q}(t,x) =& \nabla^2 q(t, K(t)^{-1}\sqrt{C}x). \end{aligned}$$

Notice that

$$\tilde{s}(t,x) = C\nabla q(t,x), \quad \nabla\tilde{s}(t,x) = C\nabla^2 q(t,x).$$

Hence,

$$\begin{aligned} |\tilde{s}(t,\cdot)|_\infty =& |C\nabla q(t,\cdot)|_\infty = |K(t)\sqrt{C}\nabla\bar{q}(t,\cdot)|_\infty \le e^{-\frac{t}{2}}\|e^{\frac{t}{2}}K(t)\|\|\sqrt{C}\||\nabla\bar{q}(t,\cdot)|_\infty, \\ \|\nabla\tilde{s}(t,\cdot)\|_\infty =& \|C\nabla^2 q(t,\cdot)\|_\infty = \|K(t)^2\nabla^2\bar{q}(t,\cdot)\|_\infty \le e^{-t}\|e^{\frac{t}{2}}K(t)\|^2\|\nabla^2\bar{q}(t,\cdot)\|_\infty. \end{aligned}$$

Define the constant,

$$K := \sup_{t\ge 0}\|e^{\frac{t}{2}}K(t)\| = \sup_{t\ge 0}\|A\bar{A}_t^{-1}\| = \max\{1, \|AC^{-1}\|\}.$$

By Theorem 3.1 and the initial value $\bar{q}(0,x) = h(\sqrt{C}x)$, we obtain,

$$\begin{aligned} |\tilde{s}(t,\cdot)| \le& e^{-\frac{t}{2}}K\|C\|^{\frac{1}{2}}|\sqrt{C}\nabla h|_\infty, \\ \|\nabla\tilde{s}(t,\cdot)\|_\infty \le& e^{-t}K^2\Big(\|C\nabla h\|_\infty + |\sqrt{C}\nabla h|_\infty^2\Big). \end{aligned} \tag{28}$$

Let,

$$\begin{aligned} L_0 :=& K^2\Big(\|C\nabla h\|_\infty + |\sqrt{C}\nabla h|_\infty^2\Big), \\ L_1 :=& K\|C\|^{\frac{1}{2}}|\sqrt{C}\nabla h|_\infty, \end{aligned}$$

then we get the results in Corollary 3.2.

Now we turn to proof of Corollary 3.8. Recall that, in Corollary 3.8, we have,

$$\begin{cases} C = I_d, \\ A = (1-e^{-\delta})I_d, \\ \bar{A}_t = (1-e^{-t})I_d. \end{cases}$$

And the distribution of the forward OU process at time δ is given by

$$\overrightarrow{P}_\delta = \mathcal{N}\big(0, (1-e^{-\delta})I_d\big) * P_{0,\delta},$$

where $P_{0,\delta} := Law(e^{-\frac{\delta}{2}}X_0)$, $\quad X_0 \sim \overrightarrow{P}_0$ and $P_{0,\delta}$ satisfies $\mathrm{Diam}(\mathrm{Supp}\, P_{0,\delta}) \le e^{-\frac{\delta}{2}}R$. Then, by Theorem 3.7, the corresponding $h(x) = \log p_\delta(x) + \frac{|x|^2}{2(1-e^\delta)}$ satisfies,

$$|\nabla h|_\infty \le \frac{Re^{-\frac{\delta}{2}}}{1-e^{-\delta}}, \quad \|\nabla^2 h\| \le 2\frac{R^2e^{-\delta}}{(1-e^{-\delta})^2}.$$

The constants then can be computed directly,

$$\begin{cases} K = \sup_{t\geq\delta} \|A\bar{A}_t^{-1}\| = 1, \\ L_0 = K^2(\|\nabla h\|_\infty + |\nabla h|_\infty^2) \leq 3\frac{R^2 e^{-\delta}}{(1-e^{-\delta})^2}, \\ L_1 = K|\nabla h|_\infty \leq \frac{Re^{-\frac{\delta}{2}}}{1-e^{-\delta}}. \end{cases}$$

Then, by Corollary 3.2, for all $t \geq \delta$,

$$\begin{aligned} \|\nabla^2 \log p_t(x) + \bar{A}_t^{-1}\|_\infty \leq & L_0 e^{-(t-\delta)} = 3\frac{R^2}{(1-e^{-\delta})^2} e^{-t}, \\ |\nabla^2 \log p_t(x) + \bar{A}_t^{-1} x|_\infty \leq & L_1 e^{-\frac{t-\delta}{2}} = \frac{R}{1-e^{-\delta}} e^{-\frac{t}{2}}. \end{aligned} \tag{29}$$

□

A.3 Proof of Theorem 3.3

We first derive some estimates that will be useful in the proof. Since

$$\begin{aligned} \sup_{t\geq 0} \|e^t(I - C\bar{A}_t^{-1})\| = & \sup_{t\geq 0} \|(A-C)\Big(Ae^{-t} + C(1-e^{-t})\Big)^{-1}\| \\ = & \max\{\|I - CA^{-1}\|, \|AC^{-1} - I\|\}, \\ = & L_2. \end{aligned}$$

We have the following estimate,

$$\|I - C\bar{A}_t^{-1}\| \leq L_2 e^{-t}, \quad \forall t \geq 0.$$

Recall from Corollary 3.2 that, for all $t \geq 0$,

$$\|\nabla \tilde{s}(t,\cdot)\|_\infty \leq L_0 e^{-t}, \quad |\tilde{s}(t,\cdot)|_\infty \leq L_1 e^{-\frac{t}{2}},$$

then,

$$\|\nabla s(t,\cdot) + I\|_\infty \leq \|\nabla \tilde{s}(t,\cdot)\|_\infty + \|I - C\bar{A}_t^{-1}\| \leq (L_0 + L_2)e^{-t}. \tag{30}$$

Similarly, under Assumption 4, we have,

$$\forall k \in \{0, 1, ..., N-1\}, \quad \|\nabla \hat{s}_\theta(T - t_k, \cdot)\|_\infty = \|\nabla s_\theta(T - t_k, \cdot) + I\|_\infty \leq (L_0 + L_2)e^{-T+t_k}. \tag{31}$$

Throughout this proof, we set the step size

$$\tau := \sup_k (t_{k+1} - t_k) \leq \min\Big\{1, \frac{1}{(L_0 + L_2)e}\Big\},$$

and can validate that

$$\forall r > 0, \quad \frac{e^{r\tau} - 1}{r} \leq e^r \tau, \quad \text{and} \quad (L_0 + L_2)e\tau \leq 1. \tag{32}$$

Here, we present a lemma related to convergence to equilibrium for the OU process.

Lemma A.1 (Theorem 23.26 (Villani et al., 2009))**.** *Let V be λ-uniformly convex C^2 potential. Consider the Langevin process,*

$$dX_t = -\frac{1}{2}\nabla V(X_t)dt + \sqrt{C}dW_t,$$

with two initial measure μ_0 and ν_0.

$$\mathcal{W}_2(\mu_t, \nu_t) \leq \mathcal{W}_2(\mu_0, \nu_0)e^{-\frac{\lambda t}{2}}.$$

The convergence of OU is a direct consequence, with $\lambda = 1$. We also need the following martingale property for the score function.

Lemma A.2. *The quantity* $e^{-\frac{t}{2}}\left(s(T-t,\overleftarrow{X}_t)+\overleftarrow{X}_t\right)$ *is a martingale, moreover,*

$$de^{-\frac{t}{2}}\left(s(T-t,\overleftarrow{X}_t)+\overleftarrow{X}_t\right) = e^{-\frac{t}{2}}\left(I+\nabla s(T-t,\overleftarrow{X}_t)\right)\sqrt{C}d\tilde{B}_t. \tag{33}$$

Proof. By Itô's formula and (4),

$$\begin{aligned}
ds(T-t,\overleftarrow{X}_t) &= -\partial_t s(T-t,\overleftarrow{X}_t)dt + \nabla s(T-t,\overleftarrow{X}_t)d\overleftarrow{X}_t + \frac{1}{2}C\nabla\left(\nabla\cdot s(T-t,\overleftarrow{X}_t)\right)dt \\
=&-\partial_t s(T-t,\overleftarrow{X}_t) + \frac{1}{2}\left(\nabla s(T-t,\overleftarrow{X}_t)\right)\overleftarrow{X}_t dt + \left(\nabla s(T-t,\overleftarrow{X}_t)\right)s(T-t,\overleftarrow{X}_t)dt \\
&+\left(\nabla s(T-t,\overleftarrow{X}_t)\right)\sqrt{C}dB_t + \frac{1}{2}C\nabla\left(\nabla\cdot s(T-t,\overleftarrow{X}_t)\right)dt.
\end{aligned}$$

From the Fokker-Planck equation, $\partial_t p = \frac{1}{2}\left(\nabla\cdot(xp+C\nabla p)\right)$, we have,

$$\begin{aligned}
\partial_t s =&\partial_t C\nabla\log p = C\left(\frac{\nabla\partial_t p}{p} - \frac{\partial_t p}{p^2}\nabla p\right) \\
=&\frac{1}{2}C\left(\frac{\nabla\left(\nabla\cdot(xp+C\nabla p)\right)}{p} - \frac{\nabla\cdot(xp+C\nabla p)}{p^2}\nabla p\right) \\
=&\frac{1}{2}C\left(\frac{\nabla\left(dp+x\cdot\nabla p+\nabla\cdot p\left(C\nabla\log p\right)\right)}{p} - \frac{dp+x\cdot\nabla p+\nabla\cdot(pC\nabla\log p)}{p^2}\nabla p\right) \\
=&\frac{1}{2}C\left(d\nabla\log p + \frac{\nabla\left(x\cdot p\nabla\log p+(\nabla p)\cdot C\nabla\log p+p\nabla\cdot C\nabla\log p\right)}{p}\right. \\
&\left.-d\nabla\log p-(x\cdot\nabla\log p)\nabla\log p - \frac{(\nabla p)\cdot C\nabla\log p+p\nabla\cdot C\nabla\log p}{p}\nabla\log p\right) \\
=&\frac{1}{2}C\left(\nabla\log p+(x\cdot\nabla\log p)\nabla\log p+\left(\nabla^2\log p\right)x + \frac{\nabla\left((p\nabla\log p)\cdot C\nabla\log p\right)}{p}\right. \\
&+(\nabla\cdot C\nabla\log p)\nabla\log p+\nabla\left(\nabla\cdot C\nabla\log p\right)-(x\cdot\nabla\log p)\nabla\log p \\
&\left.-(\nabla\log p\cdot C\nabla\log p)\nabla\log p-(\nabla\cdot C\nabla\log p)\nabla\log p\right) \\
=&\frac{1}{2}C\left(\nabla\log p+\left(\nabla^2\log p\right)x+(\nabla\log p\cdot C\nabla\log p)\nabla\log p+2\left(\nabla^2\log p\right)C\nabla\log p\right. \\
&\left.+\nabla\left(\nabla\cdot C\nabla\log p\right)-(\nabla\log p\cdot C\nabla\log p)\nabla\log p\right) \\
=&\frac{1}{2}C\left(\nabla\log p+\left(\nabla^2\log p\right)x+2\left(\nabla^2\log p\right)C\nabla\log p+\nabla\left(\nabla\cdot C\nabla\log p\right)\right) \\
=&\frac{1}{2}\left(s+(\nabla s)x+2(\nabla s)s+C\nabla\left(\nabla\cdot s\right)\right).
\end{aligned}$$

Thus,

$$ds(T-t,\overleftarrow{X}_t) = -\frac{1}{2}s(T-t,\overleftarrow{X}_t)dt + \left(\nabla s(T-t,\overleftarrow{X}_t)\right)\sqrt{C}dB_t.$$

Combining with Itô's formula applied to $e^{-\frac{t}{2}}\left(s(T-t,\overleftarrow{X}_t)+\overleftarrow{X}_t\right)$, we obtain,

$$\begin{aligned}
&de^{-\frac{t}{2}}\left(s(T-t,\overleftarrow{X}_t)+\overleftarrow{X}_t\right)\\
=&de^{-\frac{t}{2}}\left(s(T-t,\overleftarrow{X}_t)\right)+de^{-\frac{t}{2}}\overleftarrow{X}_t\\
=&e^{-\frac{t}{2}}ds(T-t,\overleftarrow{X}_t)-\frac{1}{2}e^{-\frac{t}{2}}s(T-t,\overleftarrow{X}_t)dt-\frac{1}{2}e^{-\frac{t}{2}}\overleftarrow{X}_tdt+e^{-\frac{t}{2}}d\overleftarrow{X}_t\\
=&e^{-\frac{t}{2}}\nabla s(T-t,\overleftarrow{X}_t)\sqrt{C}d\tilde{B}_t-e^{-\frac{t}{2}}(\frac{1}{2}\overleftarrow{X}_t+s(T-t,\overleftarrow{X}_t))dt+e^{-\frac{t}{2}}d\overleftarrow{X}_t\\
=&e^{-\frac{t}{2}}\left(I+\nabla s(T-t,\overleftarrow{X}_t)\right)\sqrt{C}d\tilde{B}_t.
\end{aligned}$$

□

Proof of Theorem 3.3 For simplicity, we denote,

$$\hat{s}(t,x):=s(t,x)+x,\quad \hat{s}_\theta(t,x):=s_\theta(t,x)+x.$$

By Itô's formula and (4), we have, $\forall k\in\{0,1,...,N-1\}$ and $\forall t\in[t_k,t_{k+1})$,

$$de^{\frac{t}{2}}\overleftarrow{X}_t=\frac{1}{2}e^{\frac{t}{2}}\overleftarrow{X}_tdt+e^{\frac{t}{2}}d\overleftarrow{X}_t=e^{\frac{t}{2}}\hat{s}(T-t,\overleftarrow{X}_t)dt+e^{\frac{t}{2}}\sqrt{C}d\tilde{B}_t,$$

thus,

$$e^{\frac{t_{k+1}}{2}}\overleftarrow{X}_{t_{k+1}}=e^{\frac{t_k}{2}}\overleftarrow{X}_{t_k}+\int_{t_k}^{t_{k+1}}e^{\frac{t}{2}}\hat{s}(T-t,\overleftarrow{X}_t)dt+\sqrt{e^{t_{k+1}-t_k}}\tilde{z},$$

where $\tilde{z}_k\sim\mathcal{N}(0,C)$. Now, we couple two processed $\overleftarrow{X}_t$ and $\overleftarrow{Y}_t$ with the same Brownian motion, i.e., let $\bar{z}_k=\tilde{z}_k$ in the sampling scheme (6), then,

$$\begin{aligned}
e^{\frac{t_{k+1}}{2}}(\overleftarrow{X}_{t_{k+1}}-\overleftarrow{Y}_{t_{k+1}})=&e^{\frac{t_k}{2}}(\overleftarrow{X}_{t_k}-\overleftarrow{Y}_{t_k})+\int_{t_k}^{t_{k+1}}\left(e^{\frac{t}{2}}\hat{s}(T-t,\overleftarrow{X}_t)-e^{t-\frac{t_k}{2}}\hat{s}_\theta(T-t_k,\overleftarrow{Y}_{t_k})\right)dt\\
=&e^{\frac{t_k}{2}}(\overleftarrow{X}_{t_k}-\overleftarrow{Y}_{t_k})+\int_{t_k}^{t_{k+1}}e^{t}\left(e^{-\frac{t}{2}}\hat{s}(T-t,\overleftarrow{X}_t)-e^{-\frac{t_k}{2}}\hat{s}(T-t_k,\overleftarrow{X}_{t_k})\right)dt\\
&+\int_{t_k}^{t_{k+1}}e^{t-\frac{t_k}{2}}\left(\hat{s}(T-t_k,\overleftarrow{X}_{t_k})-\hat{s}_\theta(T-t_k,\overleftarrow{X}_{t_k})\right)dt\\
&+\int_{t_k}^{t_{k+1}}e^{t-\frac{t_k}{2}}\left(\hat{s}_\theta(T-t_k,\overleftarrow{X}_{t_k})-\hat{s}_\theta(T-t_k,\overleftarrow{Y}_{t_k})\right)dt,
\end{aligned}$$

Recall in Lemma A.2,

$$e^{-\frac{t}{2}}\hat{s}(T-t,\overleftarrow{X}_t)-e^{-\frac{t_k}{2}}\hat{s}(T-t_k,\overleftarrow{X}_{t_k})=\int_{t_k}^{t}e^{-\frac{s}{2}}\nabla\hat{s}(T-s,\overleftarrow{X}_t)\sqrt{C}d\tilde{B}_s,$$

thus,

$$\begin{aligned}\mathbb{E}\left|e^{\frac{t_{k+1}}{2}}(\overleftarrow{X}_{t_{k+1}}-\overleftarrow{Y}_{t_{k+1}})\right|^2 =&\mathbb{E}\Bigg|e^{\frac{t_k}{2}}(\overleftarrow{X}_{t_k}-\overleftarrow{Y}_{t_k})\\ &+\int_{t_k}^{t_{k+1}} e^{t-\frac{t_k}{2}}\left(\hat{s}_\theta(T-t_k,\overleftarrow{X}_{t_k})-\hat{s}_\theta(T-t_k,\overleftarrow{Y}_{t_k})\right)dt\\ &+\int_{t_k}^{t_{k+1}} e^{t-\frac{t_k}{2}}\left(\hat{s}(T-t_k,\overleftarrow{X}_{t_k})-\hat{s}_\theta(T-t_k,\overleftarrow{X}_{t_k})\right)dt\Bigg|^2\\ &+\mathbb{E}\left|\int_{t_k}^{t_{k+1}} e^t\int_{t_k}^{t} e^{-\frac{s}{2}}\nabla\hat{s}(T-s,\overleftarrow{X}_t)\sqrt{C}d\tilde{B}_s dt\right|^2\\ :=&I_1^k+I_2^k.\end{aligned}$$

By Cauchy's inequality,

$$\begin{aligned}I_2^k \leq&\mathbb{E}\left|\int_{t_k}^{t_{k+1}} e^{2t}dt\int_{t_k}^{t_{k+1}}|\int_{t_k}^{t} e^{-\frac{s}{2}}\nabla\hat{s}(T-s,\overleftarrow{X}_t)\sqrt{C}d\tilde{B}_s|^2dt\right|\\ =&\frac{e^{2t_{k+1}}-e^{2t_k}}{2}\int_{t_k}^{t_{k+1}}\mathbb{E}|\int_{t_k}^{t} e^{-\frac{s}{2}}\nabla\hat{s}(T-s,\overleftarrow{X}_t)\sqrt{C}d\tilde{B}_s|^2dt.\end{aligned}$$

By Itô's isometry and (30),

$$\begin{aligned}\mathbb{E}|\int_{t_k}^{t} e^{-\frac{s}{2}}\nabla\hat{s}(T-s,\overleftarrow{X}_t)\sqrt{C}d\tilde{B}_s|^2 \leq&\int_{t_k}^{t} e^{-s}(L_0+L_2)^2e^{-2T+2s}\mathrm{Tr}(C)ds\\ \leq&(L_0+L_2)^2\mathrm{Tr}(C)e^{-2T}(e^{t_{k+1}}-e^{t_k}).\end{aligned}$$

With sufficiently small τ satisfying (32), we have,

$$I_2^k \leq e^3(L_0+L_2)^2\mathrm{Tr}(C)e^{-2T+3t_k}\tau^2(t_{k+1}-t_k). \tag{34}$$

By the mean inequality, for any positive number f_k, we have

$$\begin{aligned}I_1^k \leq&(1+f_k)\mathbb{E}\left|e^{\frac{t_k}{2}}(\overleftarrow{X}_{t_k}-\overleftarrow{Y}_{t_k})+\int_{t_k}^{t_{k+1}} e^{t-\frac{t_k}{2}}\left(\hat{s}_\theta(T-t_k,\overleftarrow{X}_{t_k})-\hat{s}_\theta(T-t_k,\overleftarrow{Y}_{t_k})\right)dt\right|^2\\ &+(1+f_k^{-1})\mathbb{E}\left|\int_{t_k}^{t_{k+1}} e^{t-\frac{t_k}{2}}\left(\hat{s}(T-t_k,\overleftarrow{X}_{t_k})-\hat{s}_\theta(T-t_k,\overleftarrow{X}_{t_k})\right)dt\right|^2\end{aligned} \tag{35}$$

By (31) and the condition (32) on τ, the first term of (35) satisfies,

$$\begin{aligned}&\mathbb{E}\left|e^{\frac{t_k}{2}}(\overleftarrow{X}_{t_k}-\overleftarrow{Y}_{t_k})+\int_{t_k}^{t_{k+1}} e^{t-\frac{t_k}{2}}\left(\hat{s}_\theta(T-t_k,\overleftarrow{X}_{t_k})-\hat{s}_\theta(T-t_k,\overleftarrow{Y}_{t_k})\right)dt\right|^2\\ \leq&\mathbb{E}\left|e^{\frac{t_k}{2}}(\overleftarrow{X}_{t_k}-\overleftarrow{Y}_{t_k})+\int_{t_k}^{t_{k+1}} e^{t-\frac{t_k}{2}}(L_0+L_2)e^{-T+t_k}(\overleftarrow{X}_{t_k}-\overleftarrow{Y}_{t_k})dt\right|^2\\ \leq&\left(1+(L_0+L_2)e^{-T}\int_{t_k}^{t_{k+1}} e^t dt\right)^2\mathbb{E}|e^{\frac{t_k}{2}}(\overleftarrow{X}_{t_k}-\overleftarrow{Y}_{t_k})|^2\\ \leq&\left(1+3(L_0+L_2)e^{-T}(e^{t_{k+1}}-e^{t_k})\right)\mathbb{E}|e^{\frac{t_k}{2}}(\overleftarrow{X}_{t_k}-\overleftarrow{Y}_{t_k})|^2,\end{aligned}$$

Denoting

$$\epsilon_k := \sqrt{\mathbb{E}|s(T-t_k,\overleftarrow{X}_{t_k})-s_\theta(T-t_k,\overleftarrow{X}_{t_k})|^2},$$

the second term of (35) satisfies,

$$\begin{aligned}
&\mathbb{E}\left|\int_{t_k}^{t_{k+1}} e^{t-\frac{t_k}{2}}\left(\hat{s}(T-t_k,\overleftarrow{X}_{t_k})-\hat{s}_\theta(T-t_k,\overleftarrow{X}_{t_k})\right)dt\right|^2\\
=&(e^{t_{k+1}}-e^{t_k})^2e^{-t_k}\mathbb{E}|s(T-t_k,\overleftarrow{X}_{t_k})-s_\theta(T-t_k,\overleftarrow{X}_{t_k})|^2\\
=&(e^{t_{k+1}}-e^{t_k})^2e^{-t_k}\epsilon_k^2.
\end{aligned}$$

Now, for all $f_k>0$, we have,

$$\begin{aligned}
I_1^k\leq&(1+f_k)\left(1+3(L_0+L_2)e^{-T}(e^{t_{k+1}}-e^{t_k})\right)\mathbb{E}|e^{\frac{t_k}{2}}(\overleftarrow{X}_{t_k}-\overleftarrow{Y}_{t_k})|^2\\
&+(1+f_k^{-1})(e^{t_{k+1}}-e^{t_k})^2e^{-t_k}\epsilon_k^2.
\end{aligned}$$

Let,

$$f_k=e^{-T}(e^{t_{k+1}}-e^{t_k})\leq e^{\tau}-1\leq e,$$

then, by the condition (32) on τ,

$$\begin{aligned}
(1+f_k)(1+3(L_0+L_2)e^{-T}(e^{t_{k+1}}-e^{t_k}))\leq&1+f_k+3(L_0+L_2)e^{-T}(e^{t_{k+1}}-e^{t_k})+3f_k\\
=&1+\Big(4+3(L_0+L_2)\Big)e^{-T}(e^{t_{k+1}}-e^{t_k})\\
(1+f_k^{-1})(e^{t_{k+1}}-e^{t_k})^2e^{-t_k}\leq&(e+1)f_k^{-1}(e^{t_{k+1}}-e^{t_k})^2e^{-t_k}\\
=&(e+1)e^Te(t_{k+1}-t_k),
\end{aligned}$$

thus,

$$\begin{aligned}
I_1^k\leq&\Big(1+\big(4+3(L_0+L_2)\big)e^{-T}(e^{t_{k+1}}-e^{t_k})\Big)\mathbb{E}|e^{\frac{t_k}{2}}(\overleftarrow{X}_{t_k}-\overleftarrow{Y}_{t_k})|^2\\
&+(e+1)e^T\epsilon_k^2e(t_{k+1}-t_k).
\end{aligned}\tag{36}$$

Denoting $E(k)=\mathbb{E}|e^{t_k}(\overleftarrow{X}_{t_k}-\overleftarrow{Y}_{t_k})|^2$ and combining the estimates (34), (36) for I_2^k and I_1^k, we obtain,

$$\begin{aligned}
E(k+1)\leq&\Big(1+\big(4+3(L_0+L_2)\big)e^{-T}(e^{t_{k+1}}-e^{t_k})\Big)E(k)+(e+1)e^T\epsilon_k^2e(t_{k+1}-t_k)\\
&+e^3(L_0+L_2)^2\mathrm{Tr}(C)e^{-2T+3t_k}\tau^2(t_{k+1}-t_k)\\
\leq&\exp\Big(\big(4+3(L_0+L_2)\big)e^{-T+t_{k+1}}\Big)\Bigg(E(k)\exp\Big(-\big(4+3(L_0+L_2)\big)e^{-T+t_k}\Big)\\
&+(e+1)e^T\epsilon_k^2e(t_{k+1}-t_k)+e^3(L_0+L_2)^2\mathrm{Tr}(C)e^{-2T}\tau^2(e^{3t_{k+1}}-e^{3t_k})\Bigg),
\end{aligned}$$

where we used $\exp\Big(\big(4+3(L_0+L_2)\big)e^{-T+t_{k+1}}\Big)\geq 1$. Equivalently, we have

$$\begin{aligned}
&E(k+1)\exp\Big(-\big(4+3(L_0+L_2)\big)e^{-T+t_{k+1}}\Big)-E(k)\exp\Big(-\big(4+3(L_0+L_2)\big)e^{-T+t_k}\Big)\\
\leq&(e+1)e^T\epsilon_k^2e(t_{k+1}-t_k)+e^3(L_0+L_2)^2\mathrm{Tr}(C)e^{-2T}\tau^2(e^{3t_{k+1}}-e^{3t_k}).
\end{aligned}$$

Summing up over k and using the condition in Assumption 3, we obtain,

$$\begin{aligned}
&\mathbb{E}|e^{\frac{T}{2}}(\overleftarrow{X}_T-\overleftarrow{Y}_T)|^2\exp\Big(-4-3(L_0+L_2)\Big)-\mathbb{E}|(\overleftarrow{X}_0-\overleftarrow{Y}_0)|^2\exp\Big(-\big(4+3(L_0+L_2)\big)e^{-T}\Big)\\
\leq&(e+1)e^{T+1}\sum_{k=0}^{N-1}\epsilon_k^2(t_{k+1}-t_k)+e^3(L_0+L_2)^2\mathrm{Tr}(C)e^{-2T}\tau^2(e^{3T}-1)\\
\leq&(e+1)e^{T+1}\epsilon^2T+e^3(L_0+L_2)^2\mathrm{Tr}(C)e^{-2T}\tau^2(e^{3T}-1),
\end{aligned}$$

Denoting $K_1 := e^{4+3(L_0+L_2)}\max\{(e+1)e, e^3(L_0+L_2)^2\}$, we have the following bound for $\mathbb{E}|\overleftarrow{X}_T - \overleftarrow{Y}_T|^2$,

$$\mathbb{E}|\overleftarrow{X}_T - \overleftarrow{Y}_T|^2 \leq K_1\left(e^{-T}\mathbb{E}|\overleftarrow{X}_0 - \overleftarrow{Y}_0|^2 + \epsilon^2 T + \mathrm{Tr}(C)\tau^2\right).$$

Now we pick a ξ-optimal coupling of $\overleftarrow{X}_0 \sim \overrightarrow{P}_T$ and $\overleftarrow{Y}_0 \sim \overleftarrow{Q}_0 = \gamma_C$ in the Wasserstein distance, i.e.

$$\mathbb{E}|\overleftarrow{X}_0 - \overleftarrow{Y}_0|^2 \leq \mathcal{W}_2^2(\overrightarrow{P}_T, \overleftarrow{Q}_0) + \xi,$$

and obtain,

$$\mathcal{W}_2^2(\overrightarrow{P}_0, \overleftarrow{Q}_T) \leq \mathbb{E}|\overleftarrow{X}_T - \overleftarrow{Y}_T|^2 \leq K_1(e^{-T}(\mathcal{W}_2^2(\overrightarrow{P}_T, \overleftarrow{Q}_0) + \xi) + \epsilon^2 T + \mathrm{Tr}(C)\tau^2).$$

Since ξ is arbitrary, the bound will be,

$$\mathcal{W}_2^2(\overrightarrow{P}_0, \overleftarrow{Q}_T) \leq K_1(e^{-T}\mathcal{W}_2^2(\overrightarrow{P}_T, \overleftarrow{Q}_0) + \epsilon^2 T + \mathrm{Tr}(C)\tau^2).$$

Noticing that,

$$\mathcal{W}_2^2(\overrightarrow{P}_0, \gamma_C) \leq \mathbb{E}_{\overrightarrow{P}_0 \otimes \gamma_C}|x-y|^2 = \mathbb{E}_{\overrightarrow{P}_0}|x|^2 + \mathbb{E}_{\gamma_C}|y|^2 = M_2 + \mathrm{Tr}(C),$$

and by Lemma A.1,

$$\mathcal{W}_2^2(\overrightarrow{P}_T, \overleftarrow{Q}_T) = \mathcal{W}_2^2(\overrightarrow{P}_T, \gamma_C) \leq e^{-T}\mathcal{W}_2^2(\overrightarrow{P}_0, \gamma_C).$$

We get,

$$\mathcal{W}_2^2(\overrightarrow{P}_0, \overleftarrow{Q}_T) \leq K_1\left(e^{-2T}\left(M_2 + \mathrm{Tr}(C)\right) + \epsilon^2 T + \mathrm{Tr}(C)\tau^2\right). \tag{37}$$

Remark A.3. For the early stop technique, following the same approach, we can get,

$$\mathcal{W}_2^2(\overrightarrow{P}_\delta, \overleftarrow{Q}_{T-\delta}) \leq K_1\left(e^{-2T+\delta}\left(M_2 + \mathrm{Tr}(C)\right) + \epsilon^2 T + \mathrm{Tr}(C)\tau^2\right). \tag{38}$$

A.4 Result under the assumptions in Remark 3.6

Here we discuss the relaxation of Assumption 4 and the resulting convergence and complexity bound. The preliminary estimate (A.3) and (30) in Appendix A.3 still hold. Now we denote,

$$\beta_k := \|\nabla \tilde{s}_\theta(T - t_k, \cdot)\|_\infty.$$

Then the estimate (31) should be replaced by,

$$\forall k \in \{0, 1, ..., N-1\}, \quad \|\nabla \hat{s}_\theta(T - t_k, \cdot)\|_\infty \leq \beta_k + L_2 e^{-T+t_k}. \tag{39}$$

In this section, we set the step size

$$\tau := \sup_k (t_{k+1} - t_k) \leq \inf_k \left\{1, \frac{1}{(L_2 + \beta_k)e}\right\},$$

so that

$$\forall r > 0 \quad \frac{e^{r\tau} - 1}{r} \leq e^r \tau, \quad \text{and} \quad \forall k \in \{0, 1, .., N-1\}, \quad (L_2 + \beta_k)e\tau \leq 1. \tag{40}$$

Following the same approach in Appendix A.3, we still have, for all $k \in \{0, 1, ..N-1\}$

$$\mathbb{E}\left|e^{\frac{t_{k+1}}{2}}(\overleftarrow{X}_{t_{k+1}} - \overleftarrow{Y}_{t_{k+1}})\right|^2 = I_1^k + I_2^k,$$

and I_2^k satisfies (36). With (39) and (40), I_1^k now satisfies, for all $f_k > 0$,

$$\begin{aligned} I_1^k \leq & \Big(1 + \big(4 + 3(L_2 e^{-T} + \beta_k e^{-t_k})\big(e^{t_{k+1}} - e^{t_k}\big)\Big)\mathbb{E}|e^{\frac{t_k}{2}}(\overleftarrow{X}_{t_k} - \overleftarrow{Y}_{t_k})|^2 \\ & + (e+1)e^T \epsilon_k^2 e(t_{k+1} - t_k). \end{aligned}$$

Denote,

$$B_0 = 0 \quad \text{and} \quad B_k = \sum_{i=0}^{k-1} \beta_i (t_{i+1} - t_i), \quad k = 1, 2, ..., N-1.$$

Then,

$$\begin{aligned} I_1^k \leq & \Big(1 + \big(4 + 3L_2\big)e^{-T}\big(e^{t_{k+1}} - e^{t_k}\big) + 3e(B_{k+1} - B_k)\Big)\mathbb{E}|e^{\frac{t_k}{2}}(\overleftarrow{X}_{t_k} - \overleftarrow{Y}_{t_k})|^2 \\ & + (e+1)e^T \epsilon_k^2 e(t_{k+1} - t_k). \end{aligned}$$

And the recursion for $E(k)$ becomes,

$$\begin{aligned} & E(k+1)\exp\Big(-\big(4 + 3L_2\big)e^{-T+t_{k+1}} - 3eB_{k+1}\Big) - E(k)\exp\Big(-\big(4 + 3L_2\big)e^{-T+t_k} - 3eB_k\Big) \\ \leq & (e+1)e^T \epsilon_k^2 e(t_{k+1} - t_k) + e^3(L_0 + L_2)^2 \mathrm{Tr}(C) e^{-2T} \tau^2 (e^{3t_{k+1}} - e^{3t_k}). \end{aligned}$$

Summing up over k, the bound for $\mathbb{E}|\overleftarrow{X}_T - \overleftarrow{Y}_T|^2$ is of the same form as the one in Appendix A.3, with the different constant coefficient,

$$\mathbb{E}|\overleftarrow{X}_T - \overleftarrow{Y}_T|^2 \leq K_1' \left(e^{-T}\mathbb{E}|\overleftarrow{X}_0 - \overleftarrow{Y}_0|^2 + \epsilon^2 T + \mathrm{Tr}(C)\tau^2\right),$$

where $K_1' := e^{4+3(L_2+eB)} \max\{(e+1)e, e^3(L_0 + L_2)^2\}$. The remaining proof follows the same approach as in Appendix A.3, and ultimately yields,

$$\mathcal{W}_2^2(\overrightarrow{P}_0, \overleftarrow{Q}_T) \leq K_1' \left(e^{-2T}\left(M_2 + \mathrm{Tr}(C)\right) + \epsilon^2 T + \mathrm{Tr}(C)\tau^2\right).$$

A.5 Proof of Theorem 3.7

Proof. Recall that,

$$\begin{aligned} q_\sigma(x) &= \int_{\mathbb{R}^d} \exp(-\frac{|x-y|^2}{2\sigma^2}) Q_0(dy) = \int_{B(0,R)} \exp(-\frac{|x-y|^2}{2\sigma^2}) Q_0(dy), \\ g(x) &= \log q_\sigma(x) + \frac{|x|^2}{2\sigma^2}. \end{aligned}$$

Fixing x, direct computation shows,

$$\begin{aligned} \nabla g(x) =& \nabla \log q_\sigma(x) + \frac{x}{\sigma^2} = \frac{\int_{B(0,R)} (\frac{y-x}{\sigma^2}) \exp(-\frac{|x-y|^2}{2\sigma^2}) Q_0(dy)}{\int_{B(0,R)} \exp(-\frac{|x-y|^2}{2\sigma^2}) Q_0(dy)} + \frac{x}{\sigma^2} \\ =& \frac{\int_{B(0,R)} (\frac{y}{\sigma^2}) \exp(-\frac{|x-y|^2}{2\sigma^2}) Q_0(dy)}{\int_{B(0,R)} \exp(-\frac{|x-y|^2}{2\sigma^2}) Q_0(dy)}. \end{aligned}$$

Taking absolute value, we get,

$$|\nabla g(x)| = |\nabla \log q_\sigma(x) + \frac{x}{\sigma^2}| \leq \frac{R}{\sigma^2}.$$

For any unit direction z,

$$
\begin{aligned}
&\nabla_z \cdot \nabla_z (\log q_\sigma + \frac{|x|^2}{2\sigma^2}) \\
=&\frac{1}{\sigma^2}\nabla_z\Big(\frac{\int_{B(0,R)}(y\cdot z)\exp(-\frac{|x-y|^2}{2\sigma^2})Q_0(dy)}{\int_{B(0,R)}\exp(-\frac{|x-y|^2}{2\sigma^2})Q_0(dy)}\Big) \\
=&\frac{1}{\sigma^2}\Big(\frac{\int_{B(0,R)}(y\cdot z)(-\frac{x-y}{\sigma^2}\cdot z)\exp(-\frac{|x-y|^2}{2\sigma^2})Q_0(dy)}{\int_{B(0,R)}\exp(-\frac{|x-y|^2}{2\sigma^2})Q_0(dy)} \\
&-\frac{\int_{B(0,R)}(y\cdot z)\exp(-\frac{|x-y|^2}{2\sigma^2})Q_0(dy)\int_{B(0,R)}(-\frac{x-y}{\sigma^2}\cdot z)\exp(-\frac{|x-y|^2}{2\sigma^2})Q_0(dy)}{(\int_{B(0,R)}\exp(-\frac{|x-y|^2}{2\sigma^2})Q_0(dy))^2}\Big) \\
=&\frac{1}{\sigma^4}\Big(\frac{\int_{B(0,R)}(y\cdot z)^2\exp(-\frac{|x-y|^2}{2\sigma^2})Q_0(dy)}{\int_{B(0,R)}\exp(-\frac{|x-y|^2}{2\sigma^2})Q_0(dy)}-\frac{(\int_{B(0,R)}(y\cdot z)\exp(-\frac{|x-y|^2}{2\sigma^2})Q_0(dy))^2}{(\int_{B(0,R)}\exp(-\frac{|x-y|^2}{2\sigma^2})Q_0(dy))^2}\Big).
\end{aligned}
$$

Taking absolute value again, we obtain,

$$
|\nabla_z \cdot \nabla_z g(x)| = |\nabla_z \cdot \nabla_z (\log q_\sigma + \frac{|x|^2}{2\sigma^2})| \le \frac{2R^2}{\sigma^4}.
$$

Thus,

$$
\|\nabla^2 g\|_\infty \le \frac{2R^2}{\sigma^4}.
$$

□

A.6 Proof of Theorem 3.9

Proof. The proof follows the same procedure as in A.3; it suffices to substitute the corresponding values of the constants. By Corollary 3.8,

$$
\begin{cases}
C = I_d, \\
L_0 = 3\frac{R^2}{(1-e^{-\delta})^2}, \\
L_2 = \sup_{t\ge\delta}\|e^t(I - C\bar{A}_t^{-1})\| = \frac{1}{1-e^{-\delta}}.
\end{cases}
$$

By Remark A.3 and noticing that under bounded-support assumption,

$$
M_2 = \mathbb{E}|\overrightarrow{X}_0|^2 \le R^2,
$$

we get,

$$
\begin{aligned}
\mathcal{W}_2(\overrightarrow{P}_\delta, \overleftarrow{Q}_{T-\delta}) \le& e^{4+3(L_0+L_2)}e^3(L_0+L_2)^2\left(e^{-2T+\delta}\left(R^2+d\right)+\epsilon^2 T + d\tau^2\right) \\
\le& 2e^{7+4(L_0+L_2)}\left(e^{-2T+\delta}\left(R^2+d\right)+\epsilon^2 T + d\tau^2\right) \\
=& 2\exp\left(7+\frac{12R^2}{(1-e^{-\delta})^2}+\frac{4}{1-e^{-\delta}}\right)\left(e^{-2T+\delta}\left(R^2+d\right)+\epsilon^2 T + d\tau^2\right)
\end{aligned}
$$

□

A.7 Proof of Theorem 3.12

Proof. We first validate that the posterior follows Gaussian tail Assumption 2 with $A = C$, $h(x) = -\frac{|G(x)-y|_\Sigma^2}{2}$, and,

$$
\begin{aligned}
|\sqrt{C}\nabla h(x)| =& |\sqrt{C}\nabla G(x)\Sigma^{-1}(G(x)-y)| \le \|C\|^{\frac{1}{2}}(|G|_\infty + |y|)\|\Sigma\|^{-1}\|\nabla G\|_\infty, \\
\|C\nabla^2 h(x)\| =& \|C\nabla^2 G(x)\Sigma^{-1}(G(x)-y) + C\nabla G(x)\Sigma^{-1}\nabla G(x)^T\| \\
\le& \|C\|\|\Sigma\|^{-1}\Big(\|\nabla^2 G\|_\infty(|G|_\infty + |y|) + \|\nabla G\|_\infty^2\Big).
\end{aligned}
$$

Then, it suffices to substitute the corresponding values of the constants in Appendix A.3,

$$\begin{aligned}
K =& \max\{1, \|AC^{-1}\|\} = 1,\\
L_0 =& K^2\big(\|C\nabla^2 h\|_\infty + |\sqrt{C}\nabla h|_\infty^2\big)\\
\leq& \|C\|\|\Sigma\|^{-1}\Bigg(\Big(\|\nabla^2 G\|_\infty(|G|_\infty + |y|) + \|\nabla G\|_\infty^2\Big) + \|\Sigma\|^{-1}\|\nabla G\|_\infty^2\Big(|G|_\infty + y\Big)^2\Bigg),\\
L_2 =& \max\{\|I - CA^{-1}\|, \|AC^{-1} - I\|\} = 0.
\end{aligned}$$

Now following the proof in Appendix A.3, we have,

$$\mathcal{W}_2^2\big(\overrightarrow{P}_0(\cdot; y), \overleftarrow{Q}_T(\cdot; y)\big) \leq K_3\Big(e^{-2T}\big(M_2 + \mathrm{Tr}(C)\big) + \epsilon^2 T + \mathrm{Tr}(C)\tau^2\Big),$$

where

$$\begin{aligned}
K_3 =& e^{4+3k_3}\max\{(e+1)e, e^3 k_3^2\},\\
k_3 =& \|C\|\|\Sigma\|^{-1}\Bigg(\Big(\|\nabla^2 G\|_\infty(|G|_\infty + |y|) + \|\nabla G\|_\infty^2\Big) + \|\Sigma\|^{-1}\|\nabla G\|_\infty^2\Big(|G|_\infty + y\Big)^2\Bigg).
\end{aligned}$$

□

B Theories towards the generative diffusion model in infinite dimension

B.1 Motivating example towards the infinite dimensional result

We consider the following target distribution,

$$p_0(x) = \frac{1}{Z}\prod_{i=1}^{d}\Big(\frac{1}{2}\exp(-\frac{|x-\sqrt{C_i}|^2}{2C_i}) + \frac{1}{2}\exp(-\frac{|x+\sqrt{C_i}|^2}{2C_i})\Big),$$

where C is an $d \times d$ matrix diagonal matrix with $\{C_i\}_i$ as diagonal entries. Denote the case $d = 1$ and $C_1 = 1$ as,

$$p_0^1(x) = \frac{1}{2\sqrt{2\pi}}\Big(\exp(-\frac{(x-1)^2}{2}) + \exp(-\frac{(x+1)^2}{2})\Big),$$

and if one considers to apply the forward process (1), one can get the distribution at time t,

$$p_t^1(x) = \frac{1}{2\sqrt{2\pi}}\Big(\exp(-\frac{(x-e^{-\frac{t}{2}})^2}{2}) + \exp(-\frac{(x+e^{-\frac{t}{2}})^2}{2})\Big).$$

Simple calculation shows,

$$\mathcal{W}_2\big(p_0, \mathcal{N}(0, C)\big) \leq \sqrt{2\mathrm{Tr}(C)}, \quad KL\big(p_0||\mathcal{N}(0, C)\big) = d \cdot KL\big(p_0^1||\mathcal{N}(0, 1)\big). \tag{41}$$

And if one considers applying the forward process (1), one can also show that,

$$\mathcal{W}_2\big(p_t, N(0, C)\big) \leq e^{-\frac{t}{2}}\sqrt{2\mathrm{Tr}(C)}, \quad KL\big(p_t||\mathcal{N}(0, C)\big) = d \cdot KL\big(p_t^1||\mathcal{N}(0, 1)\big). \tag{42}$$

From (41) and (42), we observe that *when increasing the dimension of C while keeping* $\mathrm{Tr}(C)$ *fixed,* the Wasserstein-2 bounds only scale with the trace of C while the KL bounds scale with d.

B.2 Defining diffusion model in infinite dimension

Now we consider a separable Hilbert space H with inner product $\langle \cdot, \cdot \rangle_H$. The forward process in H has the same form as (1),

$$d\overrightarrow{X}_t = -\frac{1}{2}\overrightarrow{X}_t dt + \sqrt{C} dB_t, \quad 0 \leq t \leq T, \tag{43}$$

where C becomes a degenerate positive trace-class. We still denote the marginal distributions of $\overrightarrow{X}_t$ by $\overrightarrow{P}_t$, which converges to the stationary distribution $\mathcal{N}(0, C)$ as $t \to \infty$ (Da Prato & Zabczyk (2014), Theorem 11.11). After time reversal $\overleftarrow{X}_t = \overrightarrow{X}_{T-t}$, $\overleftarrow{X}_t$ satisfies the backward SDE (Pidstrigach et al. (2023b), Theorem 1),

$$d\overleftarrow{X}_t = \big(\frac{1}{2}\overleftarrow{X}_t + s(T-t, \overleftarrow{X}_t)\big)dt + \sqrt{C} d\tilde{B}_t, \tag{44}$$

where score function $s(t, x)$ is defined as:

$$s(t, x) := -\frac{1}{1 - e^{-t}} \mathbb{E}[\overrightarrow{X}_t - e^{-\frac{t}{2}} \overrightarrow{X}_0 | \overrightarrow{X}_t = x], \tag{45}$$

which is almost surely continuous in t with respect to the norm $\|\|_H$ and equal to $C\nabla \log p_t(x)$ when $H = \mathbb{R}^d$. We consider $s_\theta(t, x)$ to approximate the score operator in H (45) and the scheme (6) for the sampling process. We now introduce the infinite-dimensional Gaussian tail assumption.

Assumption 6. *The initial data distribution $\overrightarrow{P}_0$ has finite second moment M_2 and has a Gaussian tail, i.e.,*

$$d\overrightarrow{P}_0(x) \propto \exp\big(h(x)\big) d\mathcal{N}(0, A)(x), \tag{46}$$

where A is a degenerate positive trace-class operator that is simultaneously diagonalizable with C, A and C correspond to the same Cameron–Martin space, and both AC^{-1} and $A^{-1}C$ are bounded linear maps. The function h is two times differentiable and satisfies,

$$\begin{aligned} \|\sqrt{C}\nabla h\|_{H,\infty} &:= \sup_x \|\nabla \sqrt{C} h(x)\|_H < \infty, \\ \|C\nabla^2 h\|_{\mathcal{L}(H),\infty} &:= \sup_x \|C\nabla^2 h(x)\|_{\mathcal{L}(H)} < \infty. \end{aligned} \tag{47}$$

To extend our Corollary 3.2 and Theorem 3.3 to the infinite-dimensional case, we first follow the approach in Appendix E of (Pidstrigach et al., 2023a) that projects H onto a finite-dimensional subspace H^D, and approximates the infinite-dimensional case via the results established on H^D.

Suppose C (A, correspondingly) has an orthonormal basis e_i of eigenvectors and corresponding non-negative eigenvalues $c_i \geq 0$ ($a_i \geq 0$, correspondingly), we define the linear span of the first D eigenvectors as

$$H^D = \text{span}\{e_1, e_2, \ldots, e_D\}.$$

Let $P^D : H \to H^D$ be the orthogonal projection onto H^D. We define the finite-dimensional approximations of μ_{data} by $\mu^D_{data} = {P^D}_\#(\mu_{data})$ and discretize the forward process by $\overrightarrow{X}^D_t = P^D \overrightarrow{X}_t$, then $(\overrightarrow{X}^D_t)_{t\geq 0}$ will satisfy,

$$d\overrightarrow{X}^D_t = -\frac{1}{2}\overrightarrow{X}^D_t dt + \sqrt{P^D C P^D} dB_t. \tag{48}$$

We denote the marginal distribution of $\overrightarrow{X}^D_t$ by $\overrightarrow{P}^D_t$, and its density by p^D_t. The corresponding backward process will be,

$$d\overleftarrow{X}^D_t = (\frac{1}{2}\overleftarrow{X}^D_t + s^D(T-t, \overleftarrow{X}^D_t))dt + \sqrt{P^D C P^D} d\tilde{B}_t, \tag{49}$$

where $s^D(t, x) = P^D \mathbb{E}[s(t, \overrightarrow{X}_t) | P^D \overrightarrow{X}_t = x] = C^D \nabla \log p^D_t(x)$.

Then we show the projected distribution still follows the Gaussian tail assumption in finite dimensions, summarized as follows.

Lemma B.1. *The projected initial distribution $\overrightarrow{P_t^D}$ satisfies the finite-dimensional Gaussian tail Assumption 2,*

$$d\overrightarrow{P_0^D}(x^D) \propto \exp\big(h^D(x^D)\big)d\mathcal{N}(0, A^D)(x^D),$$

where

$$\begin{aligned}&|\sqrt{C^D}\nabla h^D|_\infty \le \|\sqrt{C}\nabla h\|_{H,\infty},\\ &\|C^D\nabla^2 h^D\|_\infty \le \|C\nabla^2 h\|^2_{\mathcal{L}(H),\infty} + \|\sqrt{C}\nabla h\|^2_{H,\infty}.\end{aligned} \tag{50}$$

Proof. The term $\exp(h^D(x^D))$ is given by,

$$\begin{aligned}\exp(h^D(x^D)) &= \int \exp\big(h(x^D, x^{D+1:\infty})\big)d\mathcal{N}(0, A^{D+1:\infty})(x^{D+1:\infty})\\ &= \mathbb{E}_{\mathcal{N}(0,A)}[\exp\big(h(X)\big)|X^D = x^D].\end{aligned}$$

Then, the gradient is given by,

$$\begin{aligned}\nabla h^D(x^D) &= \frac{\int \nabla_{x^D} h(x^D, x^{D+1:\infty})\exp\big(h(x^D, x^{D+1:\infty})\big)d\mathcal{N}(0, A^{D+1:\infty})(x^{D+1:\infty})}{\int \exp\big(h(x^D, x^{D+1:\infty})\big)d\mathcal{N}(0, A^{D+1:\infty})(x^{D+1:\infty})}\\ &= \mathbb{E}_{\tilde{P}_0}[\nabla_{x^D} h(X)|X^D = x^D],\end{aligned}$$

where $d\tilde{P}_0(x) \propto \exp\big(h(x)\big)d\mathcal{N}(0, A)$, and the Hessian is given by,

$$\begin{aligned}\nabla^2 h^D(x^D) =& \frac{\int \nabla^2_{x^D} h(x^D, x^{D+1:\infty})\exp\big(h(x^D, x^{D+1:\infty})\big)d\mathcal{N}(0, A^{D+1:\infty})(x^{D+1:\infty})}{\int \exp\big(h(x^D, x^{D+1:\infty})\big)d\mathcal{N}(0, A^{D+1:\infty})(x^{D+1:\infty})}\\ &- \Big(\frac{\int \nabla_{x^D} h(x^D, x^{D+1:\infty})\exp\big(h(x^D, x^{D+1:\infty})\big)d\mathcal{N}(0, A^{D+1:\infty})(x^{D+1:\infty})}{\int \exp\big(h(x^D, x^{D+1:\infty})\big)d\mathcal{N}(0, A^{D+1:\infty})(x^{D+1:\infty})}\Big)^{\otimes 2}\\ =&\mathbb{E}_{\tilde{P}_0}[\nabla^2_{x^D} h(X)|X^D = x^D] - (\mathbb{E}_{\tilde{P}_0}[\nabla_{x^D} h(X)|X^D = x^D])^{\otimes 2}.\end{aligned}$$

Taking the absolute value, we get,

$$\begin{aligned}&|\sqrt{C^D}\nabla h^D(x^D)| \le |\sqrt{C^D}\nabla_{x^D} h|_\infty \le \|\sqrt{C}\nabla h\|_{H,\infty},\\ &\|C^D\nabla^2 h^D(x^D)\| \le \|C^D\nabla^2_{x^D} h\|_\infty + |\sqrt{C^D}\nabla_{x^D} h|^2_\infty \le \|C\nabla^2 h\|^2_{\mathcal{L}(H),\infty} + \|\sqrt{c}\nabla h\|^2_{H,\infty}.\end{aligned}$$

Thus,

$$\begin{aligned}&|\sqrt{C^D}\nabla h^D|_\infty \le \|\sqrt{C}\nabla h\|_{H,\infty},\\ &\|C^D\nabla^2 h^D\|_\infty \le \|C\nabla^2 h\|^2_{\mathcal{L}(H),\infty} + \|\sqrt{C}\nabla h\|^2_{H,\infty}.\end{aligned}$$

□

Define,

$$\begin{aligned}\bar{A}_t &:= Ae^{-t} + C(1 - e^{-t}),\\ \bar{A}^D_t &:= A^D e^{-t} + C^D(1 - e^{-t}),\end{aligned}$$

and

$$\begin{aligned}\tilde{s}^D(t, x) &:= s^D(t, x) + C^D(\bar{A}^D_t)^{-1}x,\\ \tilde{s}(t, x) &:= s(t, x) + C\bar{A}_t^{-1}x.\end{aligned} \tag{51}$$

Since the projected initial distribution $\overrightarrow{P_0^D}$ satisfies the finite-dimensional Gaussian tail assumption (Assumption 2), Corollary 3.2 implies that there exist constants $L^D_0, L^D_1 \ge 0$ such that,

$$\|\nabla\tilde{s}^D(t, \cdot)\|_{\mathcal{L}(H^D),\infty} \le L^D_0 e^{-t}, \quad \|\tilde{s}^D(t, \cdot)\|_{H^D,\infty} \le L^D_1 e^{-\frac{t}{2}}, \quad \forall t \ge 0.$$

Here the constant L_0^D and L_1^D is given by,

$$\begin{cases} K^D = \max\{1, \|A^D(C^D)^{-1}\|\} \le \max\{1, \|AC^{-1}\|\} := K, \\ L_0^D = (K^D)^2\big(\|C^D\nabla^2 h^D\|_\infty + |\sqrt{C^D}\nabla h^D|_\infty^2\big) \le K^2\big(\|C\nabla^2 h\|_{\mathcal{L}(H),\infty} + 2\|\sqrt{C}\nabla h\|_{H,\infty}^2\big), \\ L_1^D = K^D\|C^D\|^{1/2}|\sqrt{C^D}\nabla h^D|_\infty \le K\|C\|^{\frac{1}{2}}\|\sqrt{C}\nabla h\|_{H,\infty}. \end{cases}$$

By (Pidstrigach et al., 2023a) lemma 5 (listed as Proposition B.4), we have

$$\tilde{s}^D(t, \overrightarrow{X}_t^D) \to \tilde{s}(t, \overrightarrow{X}_t) = s(t, \overrightarrow{X}_t) + C\bar{A}_t^{-1}\overrightarrow{X}_t, \quad \text{both } a.s. \text{ and } \ L_2, \quad \text{as } D \to \infty. \tag{52}$$

Thus, we obtain the following regularity estimation for the modified score,

Theorem B.2. *Under the infinite dimensional Gaussian tail Assumption 6, the modified score $\tilde{s}$ defined by* (45) *and* (51) *satisfies,*

$$\|\nabla\tilde{s}(t,\cdot)\|_{\mathcal{L}(H),\infty} \le \tilde{L}_0 e^{-t}, \quad \|\tilde{s}(t,x)\|_{H,\infty} \le \tilde{L}_1 e^{-\frac{t}{2}}, \quad \forall t > 0. \tag{53}$$

where

$$\begin{cases} K = \max\{1, \|AC^{-1}\|\}, \\ \tilde{L}_0 = K^2\big(\|C\nabla^2 h\|_{\mathcal{L}(H),\infty} + 2\|\sqrt{C}\nabla h\|_{H,\infty}^2\big), \\ \tilde{L}_1 = K\|C\|^{\frac{1}{2}}\|\sqrt{C}\nabla h\|_{H,\infty}. \end{cases}$$

(53) pertains to Corollary 3.2 formulated in the infinite-dimensional case. To derive an error bound, we also need Assumption 3 and 4 under an infinite-dimensional setting, listed as follows.

Assumption 7. *For each time discretization point t_k, $0 \le k \le N-1$, the approximated score $\tilde{s}_\theta$ satisfies,*

$$\|\nabla\tilde{s}_\theta(T-t_k,\cdot)\|_{\mathcal{L}(H),\infty} \le \tilde{L}_0 e^{-T+t_k},$$
$$\frac{1}{T}\sum_{k=0}^{N-1}(t_{k+1}-t_k)\mathbb{E}\|s(T-t_k, \overrightarrow{X}_{T-t_k}) - s_\theta(T-t_k, \overrightarrow{X}_{T-t_k})\|_H^2 \le \epsilon^2.$$

To measure the distance between the samples and the target distribution, we introduce the Wasserstein-2-distance on the Hilbert space H,

$$\mathcal{W}_2(\mu,\nu) = \left(\inf_\pi \int \|x-y\|_H^2 d\pi(x,y)\right)^{1/2},$$

where π runs over all measures on $H \times H$ with marginals μ and ν.

Now we present the convergence bound aligned with the main Theorem 3.3 under the infinite-dimensional setting.

Theorem B.3. *Suppose Assumptions 6 and 7 hold under the infinite infinite-dimensional setting. Then, sampling via scheme* (6) *with sufficiently small step size $\tau := \sup_k(t_{k+1}-t_k)$ yields*

$$\mathcal{W}_2^2(\overrightarrow{P}_0, \overleftarrow{Q}_T) \le \tilde{K}_1\Big(e^{-2T}\left(M_2 + \mathrm{Tr}(C)\right) + \epsilon^2 T + \mathrm{Tr}(C)\tau^2\Big), \tag{54}$$

where $\tilde{K}_1 = e^{4+3(\tilde{L}_0+L_2)}\max\{(e+1)e, e^3(\tilde{L}_0+L_2)^2\}$.

Proof. First, we list the preliminary estimates as in Appendix A.3. Denote

$$\hat{s}(t,x) := s(t,x) + x, \quad \hat{s}_\theta(t,x) := s_\theta(t,x) + x$$

and

$$L_2 := \max\{\|I - CA^{-1}\|, \|AC^{-1} - I\|\} = \sup_t e^t\|I - C\bar{A}_t^{-1}\|.$$

By Theorem B.2 and Assumption 7, we have,

$$\forall t \geq 0, \quad \|\nabla \hat{s}(t,\cdot)\|_{\mathcal{L}(H),\infty} \leq (\tilde{L}_0 + L_2)e^{-t},$$
$$\forall k \in \{0,1,...,N-1\}, \quad \|\nabla \hat{s}_\theta(T-t_k,\cdot)\|_{\mathcal{L}(H),\infty} \leq (\tilde{L}_0 + L_2)e^{-T+t_k}.$$

We set the step size

$$\tau := \sup_k (t_{k+1} - t_k) \leq \min\Big\{1, \frac{1}{(L_0+L_2)e}\Big\},$$

so that

$$\forall r > 0, \quad \frac{e^{r\tau}-1}{r} \leq e^r \tau, \quad \text{and} \quad (L_0+L_2)e\tau \leq 1.$$

The proof proceeds in the same manner as in Appendix A.3. However, Lemma A.2 is not available here, since in the infinite-dimensional setting there is no probability density in the sense of Lebesgue measure. Therefore, we make appropriate adjustments at the points where Lemma A.2 would otherwise be used. By Theorem B.2, $\hat{s}(T-t,\overleftarrow{X}_t)$ satisfies,

$$\begin{aligned}
\mathbb{E}\|\hat{s}(T-t,\overleftarrow{X}_t)\|_H^2 =& \mathbb{E}\|\tilde{s}(T-t,\overleftarrow{X}_t) + (I - C\bar{A}_{T-t}^{-1})\overleftarrow{X}_t\|^2 \\
\leq& 2\mathbb{E}\|\tilde{s}(T-t,\overleftarrow{X}_t)\|_H^2 + 2\mathbb{E}\|(I - C\bar{A}_{T-t}^{-1})\overleftarrow{X}_t\|^2 \\
\leq& 2\tilde{L}_1^2 e^{-(T-t)} + 2L_2^2 e^{-2(T-t)}\mathbb{E}\|\overleftarrow{X}_t\|^2 \\
\leq& 2\tilde{L}_1^2 e^{-(T-t)} + 2L_2^2 e^{-2(T-t)} \max\{M_2, \mathrm{Tr}(C)\} \\
<& \infty.
\end{aligned}$$

Thus, by Proposition B.4, $\hat{s}^D(T-t,\overleftarrow{X}_t^D) := s^D(T-t,\overleftarrow{X}_t^D)$ converge to $\hat{s}(T-t,\overleftarrow{X}_t)$ in L^2. By Lemma A.2, $e^{-\frac{t}{2}}\hat{s}^D(T-t,\overleftarrow{X}_t^D)$ is a continuous time martingale. Thus, by Doob's L^2-inequality, for any $D, N \in \mathcal{N}^+$

$$\mathbb{E}[\sup_{0\leq t\leq T} |e^{-\frac{t}{2}}\hat{s}^D(T-t,\overleftarrow{X}_t^D) - e^{-\frac{t}{2}}\hat{s}^N(T-t,\overleftarrow{X}_t^N)|^2] \leq 4e^{-T}\mathbb{E}|\hat{s}^D(0,\overleftarrow{X}_T^D) - \hat{s}^N(0,\overleftarrow{X}_T^N)|^2.$$

Since the right side is Cauchy, the left side is also Cauchy, i.e., continuous time martingales $e^{-\frac{t}{2}}\{s^D(T-t,\overleftarrow{X}_t^D)\}_D$ form a Cauchy sequence under the norm,

$$\|s\| = \mathbb{E}[\sup_{t\in[0,T]} |s_t|^2].$$

The continuous martingales are closed with respect to this norm (Karatzas & Shreve (2014), Section 1.3), so $e^{-\frac{t}{2}}s(T-t,\overleftarrow{X}_t)$ is also a continuous time martingale.

Now we follow the approach in Appendix A.3. Coupling $\overleftarrow{X}_t$ and $\overleftarrow{Y}_t$ with the same Brownian motion, and using the fact that $e^{-\frac{t}{2}}s(T-t,\overleftarrow{X}_t)$ is a martingale, we obtain,

$$\mathbb{E}\left|e^{\frac{t_{k+1}}{2}}(\overleftarrow{X}_{t_{k+1}} - \overleftarrow{Y}_{t_{k+1}})\right|^2 = \tilde{I}_1^k + \tilde{I}_2^k,$$

where

$$\begin{aligned}
\tilde{I}_1^k =& \mathbb{E}\Bigg|e^{\frac{t_k}{2}}(\overleftarrow{X}_{t_k} - \overleftarrow{Y}_{t_k}) + \int_{t_k}^{t_{k+1}} e^{t-\frac{t_k}{2}}\left(\hat{s}_\theta(T-t_k,\overleftarrow{X}_{t_k}) - \hat{s}_\theta(T-t_k,\overleftarrow{Y}_{t_k})\right)dt \\
&+ \int_{t_k}^{t_{k+1}} e^{t-\frac{t_k}{2}}\left(\hat{s}(T-t_k,\overleftarrow{X}_{t_k}) - \hat{s}_\theta(T-t_k,\overleftarrow{X}_{t_k})\right)dt\Bigg|^2,
\end{aligned}$$

and

$$\tilde{I}_2^k = \mathbb{E}\left| \int_{t_k}^{t_{k+1}} e^t \left(e^{-\frac{t}{2}} \hat{s}(T-t, \overleftarrow{X}_t) - e^{-\frac{t_k}{2}} \hat{s}(T-t_k, \overleftarrow{X}_{t_k}) \right) dt \right|^2.$$

For $e^{-\frac{t}{2}}\{s^D(T-t, \overleftarrow{X}_t^D)\}_D$, using the estimate for I_2^k in Appendix A.3, we have,

$$\begin{aligned} &\mathbb{E}\left| \int_{t_k}^{t_{k+1}} e^t \left(e^{-\frac{t}{2}} \hat{s}^D(T-t, \overleftarrow{X}_t^D) - e^{-\frac{t_k}{2}} \hat{s}(T-t_k, \overleftarrow{X}_{t_k}^D) \right) dt \right|^2 \\ \leq & e^3(\tilde{L}_0 + L_2)^2 \mathrm{Tr}(C) e^{-2T+3t_k} \tau(t_{k+1} - t_k). \end{aligned}$$

Since the L^2 convergence of $e^{-\frac{t}{2}} \hat{s}^D(T-t, \overleftarrow{X}_t^D)$ to $e^{-\frac{t}{2}} \hat{s}(T-t, \overleftarrow{X}_t)$ is uniform in time, we obtain,

$$\tilde{I}_k^2 \leq e^3(\tilde{L}_0 + L_2)^2 \mathrm{Tr}(C) e^{-2T+3t_k} \tau(t_{k+1} - t_k).$$

At this stage, we have modified the parts of the proof in Appendix A.3 that rely on Lemma A.2 and obtained analogous results. The remaining steps follow directly from Appendix A.3, where the corresponding formulas naturally extend to the infinite-dimensional setting. □

Here, we list the lemma used to extend our analysis to infinite dimensions.

Proposition B.4. *(Pidstrigach et al. (2023a), Lemma 5) Let H be a separable Hilbert space and $Z, \tilde{Z}$ be two random variables taking values in H. Let e_i be an orthonormal basis of H. Denote by $H^D = \mathrm{span}\{e_1, e_2, \ldots, e_D\}$ and by P^D the projection onto H^D. Furthermore, let $Z^D = P^D \mathbb{E}[Z|P^D\tilde{Z}]$. If $\mathbb{E}\|\mathbb{E}[Z|\tilde{Z}]\|_H^2 < \infty$, then $Z^D \to \mathbb{E}[Z|\tilde{Z}]$ in L^2 and almost surely.*

C Table of constants

Table 2: Explicit expressions for the constants in the derivations

Constant	Expression
K	$\max\{1, \|AC^{-1}\|, \|A^{-1}C\|\}$
L_0	$K^2(\|C\nabla^2 h\|_\infty + \|\sqrt{C}\nabla h\|_\infty^2)$
$\tilde{L}_0$	$K^2\left(\|C\nabla^2 h\|_{\mathcal{L}(H),\infty} + 2\|\sqrt{C}\nabla h\|_{H,\infty}^2\right)$
L_1	$K\|C\|^{\frac{1}{2}}\|\sqrt{C}\nabla h\|_\infty$
$\tilde{L}_1$	$K\|C\|^{\frac{1}{2}}\|\sqrt{C}\nabla h\|_{H,\infty}$
L_2	$\max\{\|I - CA^{-1}\|, \|AC^{-1} - I\|\}$
M_0	$\max\{\mathbb{E}\|\vec{X}_0\|^2, \mathrm{Tr}C, 1\}$
M_2	$\mathbb{E}\|\vec{X}_0\|^2$
K_1	$e^{4+3(L_0+L_2)}\max\{(e+1)e, e^3(L_0+L_2)^2\}$
K_1'	$e^{4+3(L_2+eB)}\max\{(e+1)e, e^3(L_0+L_2)^2\}$
$\tilde{K}_1$	$e^{4+3(\tilde{L}_0+L_2)}\max\{(e+1)e, e^3(\tilde{L}_0+L_2)^2\}$
K_2	$2\exp\left(7 + \frac{12R^2}{(1-e^{-\delta})^2} + \frac{4}{1-e^{-\delta}}\right)$
k_3	$\|C\|\|\Sigma\|^{-1}\left(\left(\|\nabla^2 G\|_\infty(\|G\|_\infty + \|y\|) + \|\nabla G\|_\infty^2\right) + \|\Sigma\|^{-1}\|\nabla G\|_\infty^2\left(\|G\|_\infty + y\right)^2\right)$
K_3	$e^{4+3k_3}\max\{(e+1)e, e^3k_3^2\}$