Stochastic Runge-Kutta Methods: Provable Acceleration of Diffusion Models

Yuchen Wu* Yuxin Chen* Yuting Wei*

October 8, 2024

Abstract

Diffusion models play a pivotal role in contemporary generative modeling, claiming state-of-the-art performance across various domains. Despite their superior sample quality, mainstream diffusion-based stochastic samplers like DDPM often require a large number of score function evaluations, incurring considerably higher computational cost compared to single-step generators like generative adversarial networks. While several acceleration methods have been proposed in practice, the theoretical foundations for accelerating diffusion models remain underexplored. In this paper, we propose and analyze a training-free acceleration algorithm for SDE-style diffusion samplers, based on the stochastic Runge-Kutta method. The proposed sampler provably attains ε^2 error—measured in KL divergence—using $\widetilde{O}(d^{3/2}/\varepsilon)$ score function evaluations (for sufficiently small ε), strengthening the state-of-the-art guarantees $\widetilde{O}(d^3/\varepsilon)$ in terms of dimensional dependency. Numerical experiments validate the efficiency of the proposed method.

Contents

*Department of Statistics and Data Science, University of Pennsylvania; email: {wuyc14,yuxinc,ytwei}@wharton.upenn.edu.

1 Introduction

Initially introduced by Sohl-Dickstein et al. (2015) in the context of thermodynamics modeling, diffusion models now play a pivotal role in modern generative modeling, a task that aims to generate new data instances that resemble the training data in distribution. Remarkably, diffusion models are capable of producing high-quality synthetic samples, and have claimed the state-of-the-art performance across various domains, ranging from image generation (Song and Ermon, 2019; Ho et al., 2020; Song et al., 2020a; Dhariwal and Nichol, 2021; Nichol et al., 2021; Ho et al., 2022; Rombach et al., 2022; Saharia et al., 2022; Ho and Salimans, 2022), text generation (Austin et al., 2021; Li et al., 2022; Ramesh et al., 2022), speech synthesis, (Popov et al., 2021; Kim et al., 2022), time series imputation (Tashiro et al., 2021; Alcaraz and Strodthoff, 2022), reinforcement learning (Pearce et al., 2023; Hansen-Estruch et al., 2023), and molecule modeling (Anand and Achim, 2022; Xu et al., 2022; Trippe et al.). Remarkably, diffusion models have served as crucial components of mainstream content generators including Stable Diffusion (Rombach et al., 2022), DALL-E (Ramesh et al., 2022), and Imagen (Saharia et al., 2022), among others, achieving superior performance in the now rapidly growing field of generative artificial intelligence. We refer the interested reader to Yang et al. (2023) for a comprehensive survey of methods and applications pertinent to diffusion models, and to Tang and Zhao (2024); Chen et al. (2024a) for overviews of recent theoretical development.

1.1 Diffusion model overview

On a high level, diffusion models take into consideration two processes:

1) a forward process

$$X_0 \to X_1 \to \cdots \to X_K$$

that sequentially diffuses the target data distribution into an easy-to-sample prior, typically chosen as a standard Gaussian distribution;

2) a learned reverse process

$$Y_0 \to Y_1 \to \cdots \to Y_K$$

that transforms the prior (e.g., standard Gaussian) back into a distribution that resembles the target distribution, with the aim of achieving $X_k \overset{\mathrm{d}}{\approx} Y_k$ for all $k = 0, 1, \cdots, K$. A key component that enables the construction of a faithful reverse process is the estimated (Stein) score functions (Song et al., 2020b), typically represented by pre-trained neural networks. During the sampling phase, only the reverse process is implemented to generate new data instances.

Constructing the forward process is generally straightforward which often amounts to successively injecting noise into the data; in contrast, the reverse process is far more complicated, which generally involves evaluating large-scale denoising neural networks recursively (for the purpose of computing the estimated score functions) to restore the target distribution. Viewed in this light, the number of function evaluations (NFE)—more precisely, the number of times needed to compute the output of, say, denoising neural networks—oftentimes dictates the efficiency of diffusion-based samplers.

There are at least two primary approaches concerned with the construction of the reverse processes: stochastic differential equation (SDE)-based samplers, and ordinary differential equation (ODE)-based samplers (Song et al., 2020b). These samplers are based on discrete-time processes that approximate the

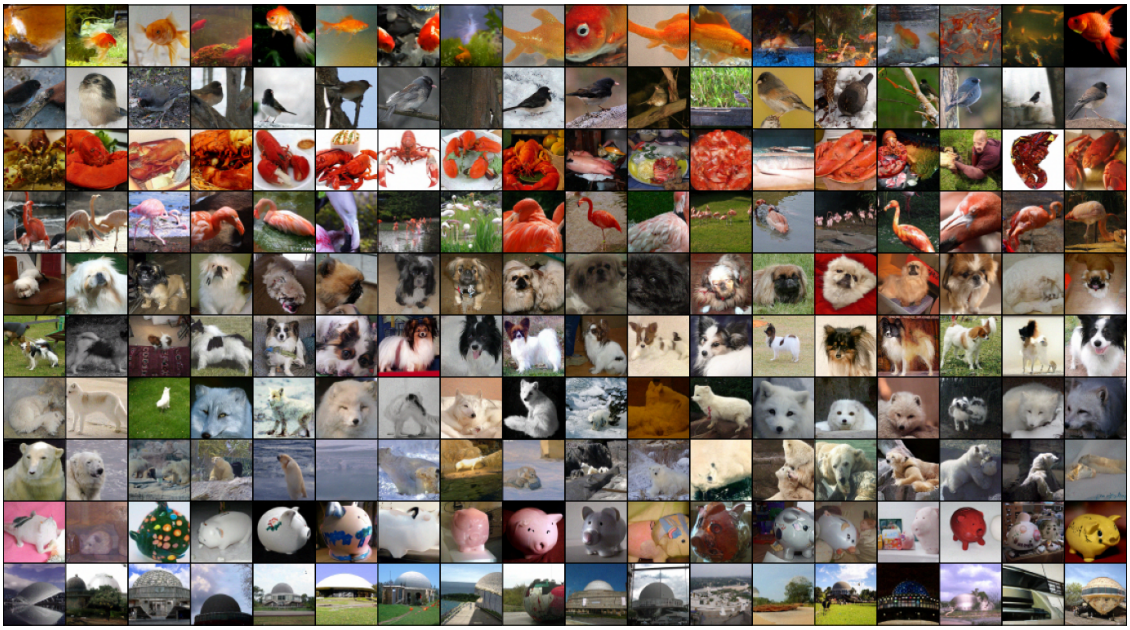


Figure 1: Class-conditional ImageNet 64×64 samples generated using 250 sampling steps with our method (Algorithm 1).

dynamics of certain diffusion SDEs and ODEs, such that when initialized at the prior, the solutions of these differential equations are designed to have marginal distributions that match the target distribution. Prominent examples of SDE-based and ODE-based samplers include the Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020) and the Denoising Diffusion Implicit Model (DDIM) (Song et al., 2020a), respectively. Empirically, ODE-based samplers offer faster sampling speeds compared to the SDE-based counterpart, while SDE-based samplers often generates higher-quality samples given sufficient runtime (Song et al., 2020a; Nichol and Dhariwal, 2021). The respective advantages of these two approaches motivate researchers to explore both types of samplers.

1.2 Accelerating diffusion models

While mainstream diffusion-based samplers like DDPM are known to generate high-fidelity samples, they often suffer from low sampling speed, requiring a large number of score function evaluations (oftentimes being neural network evaluations) to generate samples. For this reason, diffusion models incur considerably higher computational costs compared to single-step generators like generative adversarial networks (GANs) (Goodfellow et al., 2014) or variational auto-encoders (VAEs) (Kingma, 2014), thus constraining their practicality in real-world applications that demand real-time data generation.

To remedy this efficiency issue, researchers have proposed several acceleration schemes to speed up the sampling process of diffusion models. Prominent examples include the training-based method, such as model distillation (Luhman and Luhman, 2021; Salimans and Ho, 2022; Meng et al., 2023), noise level or sample trajectory learning (Nichol and Dhariwal, 2021; San-Roman et al., 2021), and consistency models (Song et al., 2023; Li et al., 2024b). Despite their impressive performance, training-based acceleration methods incur enormous additional computational costs for training, and can be challenging to implement for large-scale pre-trained diffusion models. In contrast, an alternative class of acceleration methods is based on modifying the original diffusion models without additional training, offering the flexibility to wrap around any pre-trained diffusion models (see, e.g., Lu et al. (2022b); Zheng et al. (2023); Zhao et al. (2024)). More detailed discussions about these training-free acceleration methods are deferred to Section 1.4.

Despite their empirical successes, most theoretical guarantees of diffusion acceleration are established based on ODE-based algorithms (e.g., Lee et al. (2023); Li et al. (2024a); Huang et al. (2024)). In comparison, rigorous convergence analysis for SDE-based acceleration remains largely underexplored, in spite of extensive theoretical investigation for the first-order unaccelerated solvers (Chen et al., 2023a,c; Lee et al., 2022; Benton

Sampler	Distribution	Score estimation	Complexity	Reference
SDE-based	Finite second moment	ℓ_2 score error	$\widetilde{O}(d/\varepsilon^2)$	Benton et al. (2024)
SDE-based	Bounded	ℓ_2 score error	$\widetilde{O}(d^3/\varepsilon)$	Li et al. (2024a)
SDE-based	Bounded	ℓ_2 score error	$\widetilde{O}(d^{3/2}/\varepsilon)$	This work

Table 1: The number of score function evaluations required to attain ε^2 error measured in KL divergence. In this table, we ignore the impact of score estimation errors, and focus only on SDE-based samplers. We only emphasize the dependency on d and ε, omitting logarithmic factors and other constants.

et al., 2024; Li et al., 2023; Liang et al., 2024; Li and Yan, 2024b,a). Given the popularity of stochastic samplers (Song et al., 2020b; Lu et al., 2022c; Gonzalez et al., 2024) and the fact that they tend to generate higher fidelity samples compared to their ODE-based analog, it is of great interest to design principled SDE-based acceleration schemes and demonstrate their provable advantages.

1.3 Our contributions

In this paper, we design a high-order SDE-based sampler, leveraging upon idea of stochastic Runge-Kutta methods. Our algorithm is training-free in nature. Each step only requires a single score function evaluation, introducing no extra per-step cost compared to DDPM. For a broad family of target data distributions in $\mathbb{R}^d$, it only takes $\widetilde{O}(d^{3/2}/\varepsilon)$ score function evaluations for our proposed sampler to yield a distribution that is ε^2 close to the target distribution in KL divergence, provided that the score estimates are sufficiently accurate and that ε is sufficiently small. Compared to prior theory for accelerated SDE-based samplers, our result strengthens the state-of-the-art guarantees $\widetilde{O}(d^3/\varepsilon)$ in terms of dimensional dependency. More precise comparisons between our results and previous theory on SDE-based samplers are provided in Table 1. To demonstrate the practical efficiency of the proposed method, we conduct a series of numerical experiments, as illustrated in Figure 1. More details can be found in Section 5.

1.4 Other related works

Here, we briefly discuss several other prior theory on multiple aspects of diffusion models.

Training-free acceleration schemes. A recent strand of works seeks to speed up ODE-based samplers via efficient ODE solvers. In particular, Zhang and Chen (2023) proposes DEIS, building on the semi-linear structure of the reverse process and utilizing the exponential integrator (Hochbruck and Ostermann, 2010). Similarly, Lu et al. (2022b) introduces the DPM-solver by combining high-order ODE solvers with the semi-linear framework, and further develop DPM-Solver++ to enhance stability in guided sampling (Lu et al., 2022c). Additionally, Zhao et al. (2024) establishes a predictor-corrector framework to accelerate diffusion sampling. In comparison, training-free acceleration for SDE-based samplers are considerably less explored. Jolicoeur-Martineau et al. (2021) designs an SDE solver based on stochastic Improved Euler's method. Karras et al. (2022) proposes a a stochastic sampler that comines ODE integrator with a Langevin step. Motivated by Taylor expanding diffusion processes, Lu et al. (2022c) proposes SDE-DPM-Solver++. Xue et al. (2024) presents the SA-solver, leveraging the stochastic Adams method to accelerate sampling speed. The theoretical underpinnings about these stochastic acceleration methods, however, remain far from complete.

Theory for ODE-based acceleration. In comparison to the theory for DDPM, the theoretical support for the ODE-based samplers has only been established fairly recently (Chen et al., 2023d; Li et al., 2024b; Benton et al., 2023; Chen et al., 2024b; Li et al., 2024c; Gao and Zhu, 2024; Huang et al., 2024), where the state-of-the-art convergence guarantees for the probability flow ODE are established by Li et al. (2024c). A first attempt towards the design of provably accelerated training-free ODE-based methods is made by Li et al. (2024a), which proposes and analyzes both ODE- and SDE-based acceleration algorithms. The accelerated ODE sampler proposed therein leverages a momentum-like term to enhance sample efficiency, and their accelerated SDE sampler is constructed using higher-order expansions of the conditional density. Both of

these samplers come with improved non-asymptotic convergence guarantees compared to prior theory for the unaccelerated counterpart (Benton et al., 2024; Li et al., 2024c). Furthermore, Huang et al. (2024) establish convergence guarantees for high-order ODE solvers in the context of diffusion models. Gupta et al. (2024) proposes to accelerate ODE-based samplers by incorporating a randomized midpoint method, achieving state-of-the-art dependency on the problem dimension. To bypass the complexity of developing an end-to-end theory for diffusion models, these studies often establish non-asymptotic convergence results assuming access to accurate score function estimates.

Theory for score matching/estimation. In addition to the sampling phase, the score matching phase plays a crucial role in determining the sample quality (Hyvärinen and Dayan, 2005; Lu et al., 2022a; Koehler et al., 2022). To understand the finite-sample error of score function estimation, Block et al. (2020) provides estimation guarantees under the ℓ_2 metric in terms of the Rademacher complexity of a certain concept class. Chen et al. (2023b), Oko et al. (2023) and Tang and Yang (2024) characterize the sample complexity of diffusion models when the target distribution resides within some low-dimensional linear space, the Besov space, and low-dimensional manifold, respectively. More broadly, progress has been made within the theoretical community towards addressing multiple aspects arising in score estimation (see, e.g., Oko et al. (2023); Chen et al. (2024c); Wibisono et al. (2024); Dou et al. (2024); Zhang et al. (2024); Mei and Wu (2023); Feng et al. (2024)). From a more optimization perspective, Han et al. (2024) studies the optimization error of using two-layer neural networks for score estimation.

1.5 Notation

For any positive integer n, we denote $[n] \coloneqq \{1, \ldots, n\}$. For two sequences of non-negative real numbers $\{a_n\}_{n\in\mathbb{N}_+}$ and $\{b_n\}_{n\in\mathbb{N}_+}$, we employ the notation $a_n \lesssim b_n$ (resp. $a_n \gtrsim b_n$) to indicate the existence of a universal constant C, such that $a_n \le Cb_n$ (resp. $a_n \ge Cb_n$) holds for all sufficiently large n. The notation $a_n = O(b_n)$ means $a_n \lesssim b_n$, and $\widetilde{O}(\cdot)$ hides a factor that is polynomial in $(\log d, \log \varepsilon^{-1}, \log \delta^{-1})$. For any tensor $T \in \mathbb{R}^{d_1\times d_2\times d_3}$ and matrix $M \in \mathbb{R}^{d_2\times d_3}$, we define $T[M]$ to be a vector in $\mathbb{R}^{d_1}$, such that the i-th entry of this vector is given by

$$\big(T[M]\big)_i = \sum_{j\in[d_2],k\in[d_3]} T_{ijk}M_{jk} =: \big\langle T(i,\cdot,\cdot), M\big\rangle.$$

For any vector $v \in \mathbb{R}^{d_3}$, we define Tv to be a matrix in $\mathbb{R}^{d_1\times d_2}$, such that the (i,j)-th entry of this matrix is

$$\big(Tv\big)_{i,j} = \sum_{k\in[d_3]} T_{ijk}v_k =: \big\langle T(i,j,\cdot), v\big\rangle.$$

Similarly, for any fourth order tensor $T \in \mathbb{R}^{d_1\times d_2\times d_3\times d_4}$ and third order tensor $A \in \mathbb{R}^{d_2\times d_3\times d_4}$, we define $T[A]$ to be a vector in $\mathbb{R}^{d_1}$, with the i-th entry given by

$$\big(T[A]\big)_i = \sum_{j\in[d_2],k\in[d_3],\ell\in[d_4]} T_{ijk\ell}A_{jk\ell} =: \big\langle T(i,\cdot,\cdot,\cdot), A\big\rangle.$$

For any two random objects X and Y, we say $X \perp\!\!\!\perp Y$ if and only if they are statistically independent of each other. For two distributions μ and ν, we employ $\mu\otimes\nu$ to represent the product distribution of μ and ν. For any random object X, we use $\mathcal{L}(X)$ to denote its law (i.e., distribution). Moreover, for any vector-valued function $s(t,x) : \mathbb{R}\times\mathbb{R}^d \to \mathbb{R}^d$ whose two arguments are in $\mathbb{R}$ and $\mathbb{R}^d$, respectively, we denote by $\nabla_x s(t,x) \in \mathbb{R}^{d\times d}$ (resp. $\nabla_x^2 s(t,x) \in \mathbb{R}^{d\times d\times d}$) the Jacobian matrix (resp. Hessian) w.r.t. the second argument. For any two distributions, we denote by $\mathsf{KL}(p \,\|\, q)$ the Kullback-Leibler(KL) divergence from q to p, and use $\mathsf{TV}(p,q)$ to represent the total-variation (TV) distance between p and q. For any positive integer n, we also use $\mathrm{perm}(n)$ to denote the set of permutations of $\{1,\ldots,n\}$.

2 Algorithm: a stochastic Runge-Kutta method

In this section, we present the rationale underlying the design of stochastic Runge-Kutta methods, following some preliminaries about diffusion models from an SDE perspective.

2.1 Background: diffusion models through the lens of SDEs

As mentioned previously, the diffusion generative modeling comprises a forward process and a reverse process. A widely adopted choice of the forward process can be described via the Ornstein–Uhlenbeck (OU) process

$$\mathrm{d}X_t = -X_t \mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t^{\mathrm{f}}, \qquad X_0 \sim q_0, \qquad 0 \le t \le T, \tag{1}$$

with q_0 the target data distribution, and $(B_t^{\mathrm{f}})_{0\le t\le T}$ a standard Brownian motion in $\mathbb{R}^d$ independent from X_0. As is well-known, the solution to (1) enjoys the following marginal distribution

$$X_t \overset{\mathrm{d}}{=} \lambda_t X_0 + \sigma_t Z \qquad \text{with } \lambda_t \coloneqq e^{-t} \text{ and } \sigma_t \coloneqq \sqrt{1-e^{-2t}} \tag{2}$$

for any $0 \le t \le T$, where $X_0 \sim q_0$ and $Z \sim \mathcal{N}(0, I_d)$ are independently generated. Throughout this paper, we shall denote by q_t the distribution of X_t.

How to reverse the above forward process (1) can be illuminated via a classical result in the SDE literature. To be precise, consider the following SDE

$$\mathrm{d}Y_t = \big[Y_t + 2s(t, Y_t)\big]\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad 0 \le t \le T, \tag{3}$$

where $(B_t)_{0\le t\le T}$ is also a standard Brownian motion in $\mathbb{R}^d$ independent of Y_0, and

$$s(t, x) \coloneqq \nabla_x \log q_{T-t}(x) \tag{4}$$

stands for the (Stein) score function. Classical results (Anderson, 1982; Haussmann and Pardoux, 1986) tell us that the distribution match between the above two stochastic processes in the sense that

$$Y_t \overset{\mathrm{d}}{=} X_{T-t}, \qquad 0 \le t \le T$$

as long as $Y_0 \sim q_T$, thus unveiling that (3) forms the reverse process of (1). As an important implication, if we have exact access to the score functions $\{s(t,\cdot)\}_{0\le t\le T}$ as well as q_T, then running the SDE (3) from Y_0 suffices to yield a point Y_T that exhibits the target distribution q_0.

Nevertheless, there are multiple implementation issues that prevent one from running the reverse process (3) in an exact manner. To begin with, it is unrealistic to assume exact access to the score functions; instead, one only has, for the most part, imperfect score estimates at hand. Secondly, due to the computational cost of evaluating each score function (which might involve, say, computing the output of a large neural network or transformer), it is preferable to solve the SDE (3) approximately with only a small number of score function evaluations; as a consequence, time-discretization of the SDE (3) is oftentimes necessary, in spite of the discretization error it inevitably incurs. Thirdly, the SDE (3) is typically not initialized to $Y_0 \sim q_T$, but instead, some generic data-independent distribution like $\mathcal{N}(0, I_d)$ (given that q_T can be fairly close to $\mathcal{N}(0, I_d)$ for large enough T). These issues constitute three sources that result in the discrepancy between q_0 and the distribution of Y_T, with the first two sources (i.e., the score estimation error and the discretization error) having the most significant effects.

2.2 A stochastic Runge-Kutta method

Runge-Kutta methods are a widely used family of iterative algorithms for approximating solutions to differential equations (Runge, 1895; Kutta, 1901), which enable the construction of high-order numerical solvers without requiring higher-order derivatives of the functions involved. Stochastic Runge-Kutta methods refer to a family of specialized adaptation of the general Runge-Kutta methods, designed specifically for solving SDEs (Burrage and Burrage, 1996). Motivated by the practical effectiveness of Runge-Kutta-type algorithms, we propose a high-order stochastic Runge-Kutta method—in conjunction with the use of the exponential integrator—for solving the reverse process described in Eq. (3). Here and throughout, we select K discretization time points $0 = t_0 < t_1 < \cdots < t_K < T$, and define

$$\Delta_k \coloneqq t_{k+1} - t_k \qquad \text{for } k \in \{0, 1, \ldots, K-1\}. \tag{5}$$

It is assumed that for each t_k $(0 \le k \le K)$, we only have access to the estimate $\widehat{s}(t_k, \cdot)$ of the true score function $s(t_k, \cdot) = \nabla_x \log q_{T-t_k}(\cdot)$.

Preliminaries: exponential integrator, and SDE for scores. The use of the exponential integrator arises as a common algorithmic trick in SDE to cope with linear drift components. Reformulating the SDE in Eq. (3), we obtain an equivalent form

$$\mathrm{d}\big[e^{-t}Y_t\big] = 2e^{-t}s(t,Y_t)\mathrm{d}t + \sqrt{2}\,e^{-t}\mathrm{d}B_t, \qquad 0 \le t \le T, \tag{6}$$

whose drift term does not contain a linear component as in Eq. (3). As a result, for any sequence of discretization time points $0 = t_0 < t_1 < \cdots < t_K < T$, we can take the integral to derive, for any $t \in [t_k, t_{k+1}]$,

$$Y_t = e^{t-t_k}Y_{t_k} + 2\int_0^{t-t_k} e^{t-t_k-r}s\big(t_k+r, Y_{t_k+r}\big)\mathrm{d}r + \sqrt{2}\int_0^{t-t_k} e^{t-t_k-r}\mathrm{d}B_{t_k+r}. \tag{7}$$

In addition, the evolution of $s(t, Y_t)$ can be characterized by means of another SDE. More specifically, applying the Itô formula (Øksendal, 2003) to $s(t, Y_t)$ yields

$$\mathrm{d}s(t,Y_t) = \partial_t s(t,Y_t)\mathrm{d}t + \nabla_x s(t,Y_t)\big(Y_t + 2s(t,Y_t)\big)\mathrm{d}t + \sqrt{2}\,\nabla_x s(t,Y_t)\mathrm{d}B_t + \nabla_x^2 s(t,Y_t)[I_d]\mathrm{d}t, \tag{8}$$

where we recall the notation of $\nabla_x^2 s(t,Y_t)[I_d]$ in Section 1.5.

Prelude: DDPM as a first-order Runge-Kutta solver. We now turn to designing SDE solvers through the idea of the Runge-Kutta method. Let us begin with first-order score approximation and use it to describe the first-order Runge-Kutta solver. As a natural starting point, one can approximate $s(t_k+r, Y_{t_k+r})$ in Eq. (7) by $s(t_k, Y_{t_k})$ for every $r \in [0, \Delta_k]$. With this strategy in place, we arrive at the approximation

$$Y_{t_{k+1}} \approx e^{\Delta_k}Y_{t_k} + 2(e^{\Delta_k}-1)s\big(t_k, Y_{t_k}\big) + \sqrt{2}\int_0^{\Delta_k} e^{\Delta_k-r}\mathrm{d}B_{t_k+r}$$

at the endpoint $t = t_{k+1}$. This motivates the following first-order SDE solver that computes

$$\widehat{Y}_{t_{k+1}} = e^{\Delta_k}\widehat{Y}_{t_k} + 2\big(e^{\Delta_k}-1\big)\widehat{s}\big(t_k, \widehat{Y}_{t_k}\big) + \sqrt{2}\int_0^{\Delta_k} e^{\Delta_k-r}\mathrm{d}B_{t_k+r} \tag{9}$$

iteratively for $k = 0, \ldots, K-1$, which coincides with the exponential integrator solver tailored to the DDPM sampler (Zhang and Chen, 2023).

Our algorithm: a higher-order Runge-Kutta solver. In order to further speed up the DDPM-type sampler, we seek to exploit higher-order approximation. Rearrange Eq. (7) as follows:

$$\begin{aligned} Y_{t_{k+1}} = e^{\Delta_k}Y_{t_k} + 2(e^{\Delta_k}-1)s\big(t_k, Y_{t_k}\big) + \sqrt{2}\int_0^{\Delta_k} e^{\Delta_k-r}\mathrm{d}B_{t_k+r} \\ + 2\int_0^{\Delta_k} e^{\Delta_k-r}\Big(s\big(t_k+r, Y_{t_k+r}\big) - s\big(t_k, Y_{t_k}\big)\Big)\mathrm{d}r. \end{aligned}$$

The idea is to approximate the score difference $s(t_k+r, Y_{t_k+r}) - s(t_k, Y_{t_k})$ in the last term of the above display (as opposed to approximating the score $s(t_k+r, Y_{t_k+r})$ as in the first-order solver). Let us first present the update rule of the proposed Runge-Kutta solver as follows, whose rationale will be elucidated momentarily:

$$\widehat{Y}_{t_{k+1}} = e^{\Delta_k}\widehat{Y}_{t_k} + \big(e^{\Delta_k} - e^{-\Delta_k}\big)\widehat{s}\big(t_k, \widehat{Y}_{t_k} + g_{t_k,t_{k+1}}\big) + \sqrt{2}\int_0^{\Delta_k} e^{\Delta_k-r}\mathrm{d}W_{t_k+r}, \tag{10a}$$

where $(W_t)_{0\le t\le T}$ is a standard Brownian motion in $\mathbb{R}^d$ to be used for the Runge-Kutta solver in discrete time (in order to differentiate it from the process $(B_t)_{0\le t\le T}$ used for the reverse process in Eq. (3)), and $g_{t_k,t_{k+1}}$ is a Gaussian vector defined as

$$g_{t_k,t_{k+1}} \coloneqq \frac{2\sqrt{2}}{e^{\Delta_k} - e^{-\Delta_k}}\int_0^{\Delta_k} e^{\Delta_k-r}\big(W_{t_k+r} - W_{t_k}\big)\mathrm{d}r. \tag{10b}$$

We highlight several key properties of this algorithm.

- Firstly, each iteration of (10a) only requires a single score function evaluation. This feature is in stark contrast with acceleration algorithms that demand higher-order computation.
- If we set $\alpha_k = e^{-2\Delta_k}$, then the update rule (10a) reduces to

$$\widehat{Y}_{t_{k+1}} = \frac{1}{\sqrt{\alpha_k}}\Big(\widehat{Y}_{t_k} + (1-\alpha_k)\widehat{s}(t_k, \widehat{Y}_{t_k} + \zeta_{k,1} g_{k,1})\Big) + \zeta_{k,2} g_{k,1} + \zeta_{k,3} g_{k,3}, \tag{11}$$

where $g_{k,1}, g_{k,2}, g_{k,3} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ are independent of $\widehat{Y}_{t_k}$, and $\zeta_{k,1}, \zeta_{k,2}, \zeta_{k,3} \in \mathbb{R}$ are certain functions of Δ_k defined as follows:

$$\zeta_{k,1} = \frac{2\sqrt{2} f_1(\Delta_k)^{1/2}}{e^{\Delta_k} - e^{-\Delta_k}}, \qquad \zeta_{k,2} = \frac{\sqrt{2} f_3(\Delta_k)}{f_1(\Delta_k)^{1/2}}, \qquad \zeta_{k,3} = \sqrt{2 f_2(\Delta_k) - \frac{2 f_3(\Delta_k)^2}{f_1(\Delta_k)}}, \tag{12}$$

with $f_1(\Delta) = e^{2\Delta}/2 - 2e^{\Delta} + \Delta + 3/2$, $f_2(\Delta) = e^{2\Delta}/2 - 1/2$, $f_3(\Delta) = e^{2\Delta}/2 - e^{\Delta} + 1/2$.

In addition, as $\Delta_k \to 0^+$, we have

$$\zeta_{k,1}\Delta_k^{-1/2} \to \sqrt{2/3}, \qquad \zeta_{k,2}\Delta_k^{-1/2} \to \sqrt{3/2}, \qquad \zeta_{k,3}\Delta_k^{-1/2} \to \sqrt{1/2}. \tag{13}$$

We observe that our acceleration method shares similarities with that proposed by Li et al. (2024a). They also adopt an algorithm of the form (11), but with different choices of $\zeta_{k,1}, \zeta_{k,2}, \zeta_{k,3}$: in their formulation, $\zeta_{k,1}\Delta_k^{-1/2} \to 1$, $\zeta_{k,2}\Delta_k^{-1/2} \to 1$, and $\zeta_{k,3}\Delta_k^{-1/2} \to 1$ as $\Delta_k \to 0^+$. In terms of technical motivation, their approach is based on high-order expansion of the probability density function, while our algorithm is motivated by the Runge-Kutta method for SDEs.

The whole procedure is summarized in Algorithm 1, described using the implementation-friendly form (11).

Rationale behind the construction of our Runge-Kutta solver (10). To understand the rationale underlying the above construction, we first note that in the SDE in Eq. (8), the term that dominates is $\sqrt{2}\nabla_x s(t, Y_t)\mathrm{d}B_t$, and hence it is tempting to approximate $s(t_k + r, Y_{t_k+r}) - s(t_k, Y_{t_k})$ by $\sqrt{2}\nabla_x s(t_k, Y_{t_k})(B_{t_k+r} - B_{t_k})$ to reach

$$\begin{aligned} Y_{t_{k+1}} \approx e^{\Delta_k} Y_{t_k} + 2\big(e^{\Delta_k} - 1\big) s\big(t_k, Y_{t_k}\big) + \sqrt{2}\int_0^{\Delta_k} e^{\Delta_k - r}\mathrm{d}B_{t_k+r} \\ + 2\sqrt{2}\int_0^{\Delta_k} e^{\Delta_k - r}\nabla_x s\big(t_k, Y_{t_k}\big)\big(B_{t_k+r} - B_{t_k}\big)\mathrm{d}r. \end{aligned} \tag{14}$$

Note, however, that an approach designed directly based on (14) could be computationally expensive in practice, given that it requires evaluating the Jacobian of the score function — a $(d \times d)$-dimensional object that is in general either inaccessible or too costly to estimate.

To remedy this issue, we propose an alternative solution. Observe from the Taylor expansion that

$$\begin{aligned} \big(e^{\Delta_k} - e^{-\Delta_k}\big)\Big(s\big(t_k, Y_{t_k} + g_{t_k,t_{k+1}}\big) - s\big(t_k, Y_{t_k}\big)\Big) &\approx \big(e^{\Delta_k} - e^{-\Delta_k}\big)\nabla_x s(t_k, Y_{t_k}) g_{t_k,t_{k+1}} \\ &= 2\sqrt{2}\,\nabla_x s(t_k, Y_{t_k}) \int_0^{\Delta_k} e^{\Delta_k - r}\big(B_{t_k+r} - B_{t_k}\big)\mathrm{d}r, \end{aligned} \tag{15}$$

where we take $g_{t_k,t_{k+1}} = \frac{2\sqrt{2}}{e^{\Delta_k} - e^{-\Delta_k}} \int_0^{\Delta_k} e^{\Delta_k - r}\big(B_{t_k+r} - B_{t_k}\big)\mathrm{d}r$. This suggests that the last term in Eq. (14) can be well approximated by the difference of two score functions, without the need of computing the gradient of the score functions. Substituting Eq. (15) into Eq. (14) gives

$$\begin{aligned} Y_{t_{k+1}} &\approx e^{\Delta_k} Y_{t_k} + 2\big(e^{\Delta_k} - 1\big) s\big(t_k, Y_{t_k}\big) + \sqrt{2}\int_0^{\Delta_k} e^{\Delta_k - r}\mathrm{d}B_{t_k+r} + \big(e^{\Delta_k} - e^{-\Delta_k}\big)\Big(s\big(t_k, Y_{t_k} + g_{t_k,t_{k+1}}\big) - s\big(t_k, Y_{t_k}\big)\Big) \\ &= e^{\Delta_k} Y_{t_k} + \big(e^{\Delta_k} - e^{-\Delta_k}\big) s\big(t_k, Y_{t_k} + g_{t_k,t_{k+1}}\big) + \sqrt{2}\int_0^{\Delta_k} e^{\Delta_k - r}\mathrm{d}B_{t_k+r} + \big\{2(e^{\Delta_k} - 1) - e^{\Delta_k} + e^{-\Delta_k}\big\} s\big(t_k, Y_{t_k}\big) \\ &\approx e^{\Delta_k} Y_{t_k} + \big(e^{\Delta_k} - e^{-\Delta_k}\big) s\big(t_k, Y_{t_k} + g_{t_k,t_{k+1}}\big) + \sqrt{2}\int_0^{\Delta_k} e^{\Delta_k - r}\mathrm{d}B_{t_k+r}, \end{aligned}$$

Algorithm 1: A stochastic Runge-Kutta method for diffusion models.

inputs: score estimates $\{\widehat{s}(t,\cdot)\}$, time interval $[0,T]$, discretization time points $0=t_0<\cdots<t_K<T$.
generate $Y_0\sim\mathcal{N}(0,I_d)$.
for $k=0,1,\ldots,K-1$ **do**
compute

$$\widehat{Y}_{t_{k+1}}=\frac{1}{\sqrt{\alpha_k}}\Big(\widehat{Y}_{t_k}+(1-\alpha_k)\widehat{s}\big(t_k,\widehat{Y}_{t_k}+\zeta_{k,1}g_{k,1}\big)\Big)+\zeta_{k,2}g_{k,1}+\zeta_{k,3}g_{k,3}, \tag{17}$$

where $\alpha_k=e^{-2\Delta_k}$ with $\Delta_k=t_{k+1}-t_k$, $g_{k,1},g_{k,2},g_{k,3}\sim_{i.i.d.}\mathcal{N}(0,I_d)$ are independent of $\widehat{Y}_{t_k}$, and $\zeta_{k,1},\zeta_{k,2},\zeta_{k,3}\in\mathbb{R}$ are functions of Δ_k, as defined in Eq. (12).

where the last line drops a higher-order term in view of the following approximation

$$2(e^{\Delta_k}-1)-e^{\Delta_k}+e^{-\Delta_k}=e^{\Delta_k}+e^{-\Delta_k}-2=O\big(\Delta_k^2\big). \tag{16}$$

Consequently, we arrive at the proposed approximation scheme as described in Eq. (10a). Here, we use the coefficient $(e^{\Delta_k}-e^{-\Delta_k})$ instead of $2(e^{\Delta_k}-1)$ (as suggested by the exponential integrator) for two main reasons: (1) when Δ_k is sufficiently small, the two coefficients are approximately equivalent, as shown in Eq. (16), and (2) the first coefficient is more commonly used in mainstream diffusion models (e.g., the DDPM sampler (Ho et al., 2020)).

3 Theoretical guarantees

In this section, we present a convergence theory for the proposed stochastic Runge-Kutta solver in Algorithm 1. Let us begin by stating a couple of key assumptions, with the first one concerning the boundedness of the target data distribution.

Assumption A (bounded support)**.** Assume the target distribution q_0 obeys

$$\mathbb{P}_{Y\sim q_0}\big(\|Y\|_2\le R\big)=1$$

for some quantity $R>0$. Without loss of generality, we assume throughout that $R=\sqrt{d}$. Note that for any distribution with bounded support, we can always achieve this by properly rescaling the data.

The next assumption below imposes a few conditions on the choice of the discretization time points $0=t_0<\cdots<t_K<T$, where we recall that $\Delta_k=t_{k+1}-t_k$. A concrete choice of valid discretization time points shall be provided momentarily in Corollary 3.5.

Assumption B (discretization time points)**.** Suppose that there exists $\kappa\in(0,1/4)$, such that

$$\Delta_k\le\kappa\min\{1,(T-t_{k+1})^2\},\quad k=0,1,\cdots,K-1\qquad\text{and}\qquad d^2\kappa\lesssim 1. \tag{18}$$

It is further assumed that

$$1.3\Delta_k+\big(53\Delta_k+10\Delta_k^2\big)\big(\sigma_{T-t_k}^{-2}+\lambda_{T-t_k}^2\sigma_{T-t_k}^{-4}\big)d\le 1/2,\qquad k=0,1,\cdots,K-1, \tag{19}$$

where we recall the definition of λ_t and σ_t in (2). In addition, we assume that $\delta=T-t_K>0$.

Remark 3.1. One can often interpret κ as a proxy for the step size. The upper bound $1/4$ in Assumption B is not crucial and can be replaced by any positive numerical constant.

Remark 3.2. The assumption of δ being positive implies early stopping when tracking the reverse process. This step is essential, as for non-smooth target distributions, the score function $s(t,\cdot)$ can diverge as $t\to T$. In

effect, our algorithm approximates a slightly noised distribution q_δ rather than the exact target distribution q_0, which is acceptable for a sufficiently small δ. Moreover, q_δ and q_0 are close in Wasserstein distance. This early stopping technique is commonly used in both practical applications and theoretical analysis of diffusion models (Song et al., 2020b; Benton et al., 2024).

Furthermore, we are still in need of assumptions that capture the accuracy of the estimated score functions, as stated below.

Assumption C (score estimation error)**.** Suppose that the score estimates $\{\widehat{s}(t,\cdot)\}$ satisfy the following properties:

1. For every $k = 0, 1, \cdots, K-1$, it holds that

$$\sup_{a_k \in \mathsf{I}_k, b_k \in \mathsf{I}'_k} \mathbb{E}\left[\left\|\widehat{s}(t_k, a_k Y_{t_{k+1}} + b_k g) - s(t_k, a_k Y_{t_{k+1}} + b_k g)\right\|_2^2\right] \le \varepsilon_{\mathsf{score},k}^2,$$

   where $\mathsf{I}_k = [1 - 3.1\sqrt{\Delta_k \kappa}, 1 + 3.1\sqrt{\Delta_k \kappa}]$, $\mathsf{I}'_k = [0, 3.5\Delta_k^{1/2}]$, and $Y_{t_{k+1}} \sim q_{T-t_{k+1}}$ and $g \sim \mathcal{N}(0, I_d)$ are independently generated. Recall that κ is the stepsize proxy, as defined in Assumption B.

2. For all $k = 0, 1, \cdots, K-1$, it holds that

$$\sup_{y \in \mathbb{R}^d} \frac{\sigma_{T-t_k}^2}{\lambda_{T-t_k}} \left\|\widehat{s}(t_k, y) - s(t_k, y)\right\|_2 \le 2\sqrt{d}.$$

Remark 3.3. To interpret the second point of Assumption C, observe that $\lambda_{T-t}^{-1}(\sigma_{T-t}^2 s(t,y) + y) = m(t,y)$, where $m(t,y) = \mathbb{E}[\theta \mid \lambda_{T-t}\theta + \sigma_{T-t} g = y]$ with $(\theta, g) \sim q_0 \otimes \mathcal{N}(0, I_d)$. Under Assumption A, it holds that $\|m(t,y)\|_2 \le \sqrt{d}$ for all $t \in [0,T]$ and $y \in \mathbb{R}^d$. Similarly, we define $\widehat{m}(t,y) = \lambda_{T-t}^{-1}(\sigma_{T-t}^2 \widehat{s}(t,y) + y)$. Then Assumption C is equivalent to assuming

$$\sup_{y \in \mathbb{R}^d} \left\|\widehat{m}(t,y) - m(t,y)\right\|_2 \le 2\sqrt{d}, \qquad \text{for all } t \in \{t_0, t_1, \cdots, t_{K-1}\}. \tag{20}$$

which is valid as long as $\|\widehat{m}(t,y)\|_2 \le \sqrt{d}$. For a general $\widehat{s}$, the second point of Assumption C is always satisfied by projecting $\widehat{m}(t,y) = \lambda_{T-t}^{-1}(\sigma_{T-t}^2 \widehat{s}(t,y) + y)$ onto the d-dimensional ball $\{x \in \mathbb{R}^d : \|x\|_2 \le \sqrt{d}\}$. Furthermore, this projection reduces the score estimation error. Specifically, if we denote the projection operator by P, and let $\widehat{s}^{\mathsf{P}}(t,y) := \sigma_{T-t}^{-2}(\lambda_{T-t}\mathsf{P}(\widehat{m}(t,y)) - y)$, then it holds that $\|\widehat{s}^{\mathsf{P}}(t,y) - s(t,y)\|_2 \le \|\widehat{s}(t,y) - s(t,y)\|_2$.

We are now positioned to present our convergence guarantees for the proposed Runge-Kutta method, as stated in the following theorem. Here and throughout, p_{output} stands for the distribution of the output $\widehat{Y}_{t_K}$ of Algorithm 1.

Theorem 3.4. *Under Assumptions A, B and C, Algorithm 1 achieves*

$$\mathsf{KL}(q_\delta \parallel p_{\mathsf{output}}) \lesssim d^3\kappa^2 T + d^7\kappa^3(\delta^{-1} + T) + \kappa^{1/2} d \sum_{k=0}^{K-1} \varepsilon_{\mathsf{score},k}^{1/2} \sigma_{T-t_k} \lambda_{T-t_k}^{-1/2} + d e^{-2T}.$$

The proof of Theorem 3.4 is postponed to Section 4. In the next corollary, we derive the upper bound for KL divergence based on a specific stepsize selection. The proof of this corollary is provided in Appendix C.8.

Corollary 3.5. *Consider any early stopping point $\delta \in (0, 1/2)$ and any desired accuracy level ε^2.*

1. *If $\varepsilon \le 1/\sqrt{d}$, then there exist T and $0 = t_0 < t_1 < \cdots < t_K = T - \delta$ such that Algorithm 1 achieves*

$$\mathsf{KL}(q_\delta \parallel p_{\mathsf{output}}) \lesssim \varepsilon^2 + \varepsilon^3 d^{5/2} \delta^{-1} + \sum_{k=0}^{K-1} d^{1/2} \varepsilon_{\mathsf{score},k}^{1/2}$$

   with $K = \widetilde{O}\big(d^{3/2}(\varepsilon\delta)^{-1}\big)$.

2. *If* $\sqrt{d}/2 \geq \varepsilon \geq 1/\sqrt{d}$, *then there exist* T *and* $0 = t_0 < t_1 < \cdots < t_K = T - \delta$ *such that Algorithm 1 yields*

$$\mathsf{KL}(q_\delta \,\|\, p_{\mathsf{output}}) \lesssim \varepsilon^2 + \varepsilon^3 d^{5/2}\delta^{-1} + \sum_{k=0}^{K-1} d^{1/2}\varepsilon_{\mathsf{score},k}^{1/2}$$

with $K = \widetilde{O}(d^2\delta^{-1})$.

When attaining the above upper bounds, we take $\Delta_{K-1} = \kappa\delta^2$, *and* $\Delta_{k-1} = \min\{\kappa, \Delta_k(1 + \sqrt{\kappa\Delta_k})^2\}$ *for* $k = K-1, K-2, \cdots, 1$. *More details are given in Appendix C.8.*

When the score estimation errors are negligible (i.e., $\varepsilon_{\mathsf{score},k} \approx 0$) and ε is sufficiently small, Corollary 3.5 implies that Algorithm 1 requires

$$\widetilde{O}\big(d^{3/2}(\varepsilon\delta)^{-1}\big)$$

steps—or equivalently, $\widetilde{O}(d^{3/2}(\varepsilon\delta)^{-1})$ score function evaluations—to yield a distribution that is within ε^2-KL-divergence to an early-stopped target distribution. In contrast, (1) the state-of-the-art convergence rate for unaccelerated diffusion model is $\widetilde{O}(d/\varepsilon^2)$ (Benton et al., 2024), so that our theory exhibits improved ε-dependency. Also, the iteration complexity established for the SDE-based accelerated algorithm in Li et al. (2024a) is $\widetilde{O}(d^3/\varepsilon)$, and hence our result exhibits better d-dependency than the theory presented in Li et al. (2024a). It is also noteworthy that the analyses in both Benton et al. (2024); Li et al. (2024a) offer a more favorable dependency on the score estimation error. We believe that our less favorable dependency on the score errors stems from our technical limitations as opposed to the algorithm drawback; improving this dependency calls for new techniques that we leave for future research.

Remark 3.6. It is worth noting that our theory is stated in terms of the KL divergence between the algorithm output and the target distribution. While one can simply invoke the Pinsker inequality (i.e., $\mathsf{TV}(q_\delta, p_{\mathsf{output}}) \leq \sqrt{\mathsf{KL}(q_\delta \,\|\, p_{\mathsf{output}})/2}$) to obtain an upper bound on the total-variation distance, this approach has been shown to be sub-optimal for stochastic samplers like DDPM. In fact, the concurrent work Li and Yan (2024b) has established the striking result that $\mathsf{TV}(q_\delta, p_{\mathsf{output}})$ could be order-of-magnitudes better than $\sqrt{\mathsf{KL}(q_\delta \,\|\, p_{\mathsf{output}})/2}$ for DDPM. How to obtain the desired control on the TV metric for our proposed sampler is left for future investigation.

4 Analysis

In this section, we provide an overview of the proof strategies for Theorem 3.4, with detailed proofs deferred to the appendix.

Step 1: constructing an auxiliary process. To facilitate analysis, we introduce the following auxiliary stochastic process, obtained by replacing the estimated score $\widehat{s}(t_k, \cdot)$ with the true score $s(t_k, \cdot)$ in the update rule (10a) and initializing it to the distribution q_T:

$$\begin{aligned} \overline{Y}_{t_0} &\sim q_T, \\ \overline{Y}_{t_{k+1}} &= e^{\Delta_k}\overline{Y}_{t_k} + \big(e^{\Delta_k} - e^{-\Delta_k}\big)s\big(t_k, \overline{Y}_{t_k} + g_{t_k,t_{k+1}}\big) + \sqrt{2}\int_0^{\Delta_k} e^{\Delta_k - r}\mathrm{d}W_{t_k+r}, \quad 0 \leq k < K, \end{aligned} \tag{21}$$

where we recall that $g_{t_k,t_{k+1}} \in \mathbb{R}^d$ is a Gaussian random vector defined in Eq. (10b). Given that the process (21) is defined only at the discretization time points $\{t_k\}$, we find it convenient to introduce a natural interpolation of Eq. (21) to cover all time instances: for any $t \in (t_k, t_{k+1})$, take

$$\overline{Y}_t = e^{t-t_k}\overline{Y}_{t_k} + \big(e^{t-t_k} - e^{-t+t_k}\big)s\big(t_k, \overline{Y}_{t_k} + g_{t_k,t}\big) + \sqrt{2}\int_0^{t-t_k} e^{t-t_k-r}\mathrm{d}W_{t_k+r}. \tag{22}$$

where

$$g_{t_k,t} \coloneqq \frac{2\sqrt{2}}{e^{t-t_k} - e^{-t+t_k}}\int_0^{t-t_k} e^{t-t_k-r}\big(W_{t_k+r} - W_{t_k}\big)\mathrm{d}r. \tag{23}$$

Before proceeding, let us compare the auxiliary process $(\overline{Y}_t)$ with the exact reverse process (Y_t). Recall that the true reverse process can be rearranged as (cf. Eq. (7)):

$$\begin{aligned} Y_t = e^{t-t_k} Y_{t_k} + 2(e^{t-t_k}-1)s(t_k, Y_{t_k}) + \sqrt{2}\int_0^{t-t_k} e^{t-t_k-r}\mathrm{d}B_{t_k+r} \\ + 2\int_0^{t-t_k} e^{t-t_k-r}\big(s(t_k+r, Y_{t_k+r}) - s(t_k, Y_{t_k})\big)\mathrm{d}r. \end{aligned} \tag{24}$$

Taking the differential of processes (22) and (24) and rearranging terms reveal that: for any $t \in (t_k, t_{k+1}]$,

$$\begin{aligned} \mathrm{d}\overline{Y}_t &= \Big[e^{t-t_k}\overline{Y}_{t_k} + (e^{t-t_k}+e^{-t+t_k})s\big(t_k, \overline{Y}_{t_k}+g_{t_k,t}\big) + \sqrt{2}\int_0^{t-t_k} e^{t-t_k-r}\mathrm{d}W_{t_k+r}\Big]\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t \\ &\quad + 2\sqrt{2}\,\nabla_x s\big(t_k, \overline{Y}_{t_k}+g_{t_k,t}\big)\big(W_t - W_{t_k}\big)\mathrm{d}t + 2\sqrt{2}\,\nabla_x s\big(t_k, \overline{Y}_{t_k}+g_{t_k,t}\big)\int_0^{t-t_k} e^{t-t_k-r}\big(W_{t_k+r}-W_{t_k}\big)\mathrm{d}r\mathrm{d}t \\ &\quad - \frac{2\sqrt{2}(e^{t-t_k}+e^{-t+t_k})}{e^{t-t_k}-e^{-t+t_k}}\nabla_x s\big(t_k, \overline{Y}_{t_k}+g_{t_k,t}\big)\int_0^{t-t_k} e^{t-t_k-r}\big(W_{t_k+r}-W_{t_k}\big)\mathrm{d}r\mathrm{d}t, \end{aligned} \tag{25a}$$

$$\begin{aligned} \mathrm{d}Y_t &= e^{t-t_k}\Big[Y_{t_k} + 2s(t_k, Y_{t_k}) + \sqrt{2}\int_0^{t-t_k} e^{-r}\mathrm{d}B_{t_k+r}\Big]\mathrm{d}t + 2\big(s(t, Y_t) - s(t_k, Y_{t_k})\big)\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t \\ &\quad + 2\int_0^{t-t_k} e^{t-t_k-r}\big(s(t_k+r, Y_{t_k+r}) - s(t_k, Y_{t_k})\big)\mathrm{d}r\mathrm{d}t. \end{aligned} \tag{25b}$$

For notational simplicity, we shall often abbreviate Eq. (25) by

$$\mathrm{d}\overline{Y}_t = \overline{\mathcal{F}}\Big(t, \overline{Y}_{t_k}, (W_s - W_{t_k})_{t_k\le s\le t}\Big)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t \tag{26a}$$

$$\mathrm{d}Y_t = \mathcal{F}\Big(t, (Y_s)_{t_k\le s\le t}, (B_s - B_{t_k})_{t_k\le s\le t}\Big)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \tag{26b}$$

in the sequel, where the definitions of the mappings $\mathcal{F}$ and $\overline{\mathcal{F}}$ are clear from the context. From the original definition of the reverse process in Eq. (3), it can be easily shown that (which we omit here for brevity)

$$\mathcal{F}\Big(t, (Y_s)_{t_k\le s\le t}, (B_s - B_{t_k})_{t_k\le s\le t}\Big) = Y_t + 2s(t, Y_t). \tag{27}$$

As a result, we shall often adopt the following more concise notation

$$\mathcal{F}(t, Y_t) = \mathcal{F}\Big(t, (Y_s)_{t_k\le s\le t}, (B_s - B_{t_k})_{t_k\le s\le t}\Big) = Y_t + 2s(t, Y_t). \tag{28}$$

Step 2: characterizing the effect of time-discretization errors. With the aforementioned properties in mind, this step seeks to upper bound the KL divergence between the two processes (Y_t) and $(\overline{Y}_t)$—both initialized by $Y_0 \overset{\mathrm{d}}{=} \overline{Y}_0 \sim q_T$—by means of Girsanov's Theorem. On a high level, the primary purpose of this step is to characterize the impact of the time-discretization error and the approximation error due to the application of the Runge-Kutta method, given that both the reverse process (Y_t) and its time-discretized counterpart $(\overline{Y}_t)$ are constructed using exact score functions.

To begin with, we upper bound the KL divergence between $(\overline{Y}_t)$ and (Y_t) using the differences between the mappings $\mathcal{F}$ and $\overline{\mathcal{F}}$ introduced in Step 1, as stated below.

Lemma 4.1. *Denote by $Q_{T-\delta}$ (resp. $\overline{Q}_{T-\delta}$) the distribution of $(Y_t)_{0\le t\le T-\delta}$ (resp. $(\overline{Y}_t)_{0\le t\le T-\delta}$), which we recall are respectively defined in Eqs. (24) and (22) from initialization $Y_0 \overset{\mathrm{d}}{=} \overline{Y}_0 \sim q_T$. Then under Assumption A, it holds that*

$$\mathsf{KL}(Q_{T-\delta} \,\|\, \overline{Q}_{T-\delta}) \le \sum_{k=0}^{K-1} \liminf_{\tau\to 0^+} \int_{t_k+\tau}^{t_{k+1}} \mathbb{E}\Big[\big\|\mathcal{F}(t, Y_t) - \overline{\mathcal{F}}\big(t, Y_{t_k}, (H_s^\tau)_{t_k\le s\le t}\big)\big\|_2^2\Big]\mathrm{d}t.$$

In the above display, $(Y_t)_{0\le t\le T-\delta}\sim Q_{T-\delta}$, *and* $(H_s^\tau)_{t_k\le s\le t_{k+1}}$ *is a stochastic process satisfying*

$$\begin{aligned}\mathrm{d}H_t^\tau &= \frac{1}{\sqrt{2}}\Big(\mathcal{F}(t,Y_t)-\overline{\mathcal{F}}\big(t,Y_{t_k},(H_s^\tau)_{t_k\le s\le t}\big)\Big)\mathrm{d}t+\mathrm{d}B_t, \qquad t_k+\tau\le t\le t_{k+1}\\ (H_s^\tau)_{t_k\le s\le t_k+\tau} &= (B_s-B_{t_k})_{t_k\le s\le t_k+\tau}.\end{aligned} \tag{29}$$

Note that the distribution of $(H_s^\tau)_{t_k\le s\le t}$ depends on the value of $\tau\in(0,\Delta_k)$. The proof of Lemma 4.1 is deferred to Appendix C.1. The existence and the uniqueness of process (29) are also established therein.

In the next lemma, we show the proximity of the process $(H_s^\tau)_{t_k\le s\le t}$ is and the Brownian motion increment process $(B_s-B_{t_k})_{t_k\le s\le t}$. The proof of this lemma is postponed to Appendix C.2.

Lemma 4.2. *Consider any* $\tau\in(0,\Delta_k)$. *Under Assumptions A and B, we have the following upper bounds:*

$$\begin{aligned}&\mathbb{E}\Big[\sup_{t_k\le t\le t_{k+1}}\big\|B_t-B_{t_k}-H_t^\tau\big\|_2^2\Big]\\ &\quad\lesssim \sigma_{T-t_{k+1}}^{-2}\Delta_k^4 d+(\sigma_{T-t_k}^{-2}+\lambda_{T-t_k}^2\sigma_{T-t_k}^{-4})^2\Delta_k^3 d^3+\lambda_{T-t_{k+1}}^4\sigma_{T-t_{k+1}}^{-12}\Delta_k^4 d^3+\lambda_{T-t_{k+1}}^2\sigma_{T-t_{k+1}}^{-8}\Delta_k^4 d,\\ &\mathbb{E}\Big[\sup_{t_k\le t\le t_{k+1}}\big\|B_t-B_{t_k}-H_t^\tau\big\|_2^4\Big]\\ &\quad\lesssim \sigma_{T-t_{k+1}}^{-4}\Delta_k^8 d^2+(\sigma_{T-t_k}^{-2}+\lambda_{T-t_k}^2\sigma_{T-t_k}^{-4})^4\Delta_k^6 d^6+\lambda_{T-t_{k+1}}^8\sigma_{T-t_{k+1}}^{-24}\Delta_k^8 d^6+\lambda_{T-t_{k+1}}^4\sigma_{T-t_{k+1}}^{-16}\Delta_k^8 d^2.\end{aligned}$$

With Lemmas 4.1 and 4.2 in mind, we can readily develop a more concise upper bound on $\mathsf{KL}(Q_{T-\delta}\,\|\,\overline{Q}_{T-\delta})$, as stated below. The proof of this lemma can be found in Appendix C.3.

Lemma 4.3. *Under Assumptions A and B, it holds that*

$$\mathsf{KL}(Q_{T-\delta}\,\|\,\overline{Q}_{T-\delta})\lesssim d^3\kappa^2 T+d^7\kappa^3(\delta^{-1}+T).$$

Step 3: bounding the effect of score estimation errors. We still need to take into account the impact of the score estimation error. In this regard, we recall process (10a), and denote by $\widehat{Q}_{T-\delta}^{\mathsf{dis}}$ (resp. $Q_{T-\delta}^{\mathsf{dis}}$) the distribution of $(\widehat{Y}_{t_k})_{0\le k\le K}$ (resp. $(Y_{t_k})_{0\le k\le K}$). The next lemma attempts to upper bound $\mathsf{KL}(Q_{T-\delta}^{\mathsf{dis}}\,\|\,\widehat{Q}_{T-\delta}^{\mathsf{dis}})$; its proof is deferred to Appendix C.4.

Lemma 4.4. *Suppose that Assumptions A, B and C hold, and that both processes* $(\widehat{Y}_t)$ *and* (Y_t) *are initialized to* $Y_0\overset{\mathrm{d}}{=}\widehat{Y}_0\sim q_T$. *Then, it holds that*

$$\mathsf{KL}(Q_{T-\delta}^{\mathsf{dis}}\,\|\,\widehat{Q}_{T-\delta}^{\mathsf{dis}})\lesssim d^3\kappa^2 T+d^7\kappa^3(\delta^{-1}+T)+\sum_{k=0}^{K-1}\frac{\varepsilon_{\mathsf{score},k}^{1/2}\kappa^{1/2}\sigma_{T-t_k}d}{\lambda_{T-t_k}^{1/2}}.$$

Step 4: determining the impact of initialization errors. In practice, we typically have no access to q_T, and a common strategy is to replace it with $\mathcal{N}(0,I_d)$. Consider process (10a), but with initialization $\widehat{Y}_0\sim\mathcal{N}(0,I_d)$ instead of $\widehat{Y}_0\sim q_T$. In this case, we denote the distribution of $\widehat{Y}_{T-\delta}$ by p_{output}. With $\widehat{Y}_0\sim\mathcal{N}(0,I_d)$, we denote the distribution of $(\widehat{Y}_{t_k})_{0\le k\le K}$ that follows the update rule (11) by $\widehat{P}_{T-\delta}^{\mathsf{dis}}$, in contrast to $\widehat{Q}_{T-\delta}^{\mathsf{dis}}$ which is the distribution of the same process with $\widehat{Y}_0\sim q_T$. Note that for $(y_0,y_1,\cdots,y_K)\in\mathbb{R}^{d(K+1)}$,

$$\frac{\mathrm{d}Q_{T-\delta}^{\mathsf{dis}}(y_0,\cdots,y_K)}{\mathrm{d}\widehat{P}_{T-\delta}^{\mathsf{dis}}(y_0,\cdots,y_K)}=\frac{\mathrm{d}Q_{T-\delta}^{\mathsf{dis}}(y_0,\cdots,y_K)}{\mathrm{d}\widehat{Q}_{T-\delta}^{\mathsf{dis}}(y_0,\cdots,y_K)}\cdot\frac{\mathrm{d}\widehat{Q}_{T-\delta}^{\mathsf{dis}}(y_0,\cdots,y_K)}{\mathrm{d}\widehat{P}_{T-\delta}^{\mathsf{dis}}(y_0,\cdots,y_K)}=\frac{\mathrm{d}Q_{T-\delta}^{\mathsf{dis}}(y_0,\cdots,y_K)}{\mathrm{d}\widehat{Q}_{T-\delta}^{\mathsf{dis}}(y_0,\cdots,y_K)}\cdot\frac{\mathrm{d}q_T(y_0)}{\mathrm{d}\pi_d(y_0)},$$

where π_d represents the distribution of a d-dimensional standard Gaussian random vector. Using the above distribution, we obtain

$$\mathsf{KL}(q_\delta\,\|\,p_{\mathsf{output}})\le\mathsf{KL}\big(Q_{T-\delta}^{\mathsf{dis}}\,\|\,\widehat{P}_{T-\delta}\big)=\mathsf{KL}\big(Q_{T-\delta}^{\mathsf{dis}}\,\|\,\widehat{Q}_{T-\delta}^{\mathsf{dis}}\big)+\mathsf{KL}(q_T\,\|\,\pi_d). \tag{30}$$

Further, recall that q_T has the same distribution as $\lambda_T\theta + \sigma_T g$, where $(\theta, g) \sim q_0 \otimes \mathcal{N}(0, I_d)$. Hence, the data processing inequality tells us that

$$\mathsf{KL}(q_T \parallel \pi_d) \le \mathsf{KL}\big(q_0 \otimes q_T \parallel q_0 \otimes \mathcal{N}(0, I_d)\big) = \frac{1}{2}\left(-d\log\sigma_T^2 - d + d\sigma_T^2 + e^{-2T}\mathbb{E}_{\theta\sim q_0}[\|\theta\|_2^2]\right) \lesssim de^{-2T}. \quad (31)$$

Substituting the above upper bound Eq. (31) and Lemma 4.4 into Eq. (30), we arrive at

$$\mathsf{KL}(q_\delta \parallel p_{\mathsf{output}}) \lesssim d^3\kappa^2 T + d^7\kappa^3(\delta^{-1} + T) + \sum_{k=0}^{K-1} \frac{\varepsilon_{\mathsf{score},k}^{1/2}\kappa^{1/2}\sigma_{T-t_k}d}{\lambda_{T-t_k}^{1/2}} + de^{-2T}$$

as claimed.

5 Numerical experiments

In this section, we illustrate the practical performance of Algorithm 1 on various image generation tasks. For benchmarking, we resort to the original DDPM sampling scheme (Ho et al., 2020) along with the SDE-based acceleration method proposed in Li et al. (2024a), ensuring that all methods adopt the same pre-trained score estimates.

More specifically, we utilize the pre-trained score functions from Nichol and Dhariwal (2021) and focus on two datasets: ImageNet-64 (Chrabaszcz et al., 2017) and CIFAR-10 (Krizhevsky et al., 2009). Note that we have not attempted to optimize the generative modeling performance with additional techniques (e.g., employing better score functions or training with higher quality datasets), as our primary goal is to evaluate the effectiveness of the proposed acceleration method. Our approach is compatible with a variety of diffusion model codebases and datasets, where we anticipate observing similar acceleration effects.

In our experiments, we compare the Fréchet inception distance (FID) (Heusel et al., 2017) of the images generated by the vanilla DDPM, the SDE acceleration method proposed in Li et al. (2024a), and our proposed method (the Stochastic Runge-Kutta method). FID quantifies the similarity between the distribution of the generated images and the target distribution, with lower FID values indicating greater similarity. For each method and step size combination, we generate 10^4 images and compute the corresponding FID. The numerical results are summarized below.

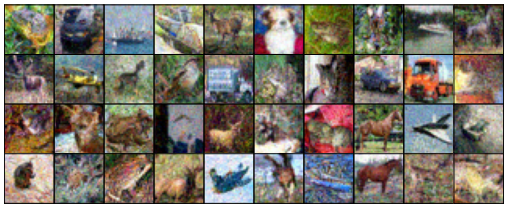
CIFAR-10. Figure 2 presents the simulation results for the CIFAR-10 dataset. The left panel shows images generated by our method and the vanilla DDPM, while the right panel illustrates the evolution of FID across different NFEs, ranging from 10 to 100. Note that here NFE is equal to the number of diffusion steps, as each step requires only one score evaluation for all methods. The generated images suggest that our method produces less noisy outputs. Moreover, our method consistently outperforms the other two methods in terms of FID across all step sizes.

ImageNet-64. Figure 3 shows the simulation results for the ImageNet-64 dataset, where we observe similar improvements as what happens for the CIFAR-10 dataset.

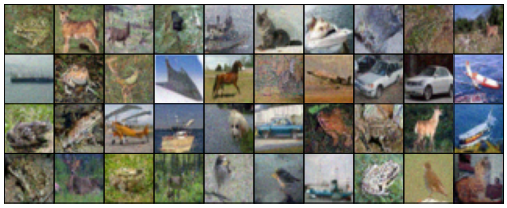
6 Discussion

In this paper, we have made progress in provably speeding up SDE-based diffusion samplers. In comparison to prior results, the convergence guarantees of our accelerated algorithm enjoy improved dimension-dependency, shedding light on the advantages of the stochastic Runge-Kutta approach. Remarkably, our algorithm paves the way for designing even higher-order SDE-based diffusion solvers, the advantages of which will be explored in future research.

Moving forward, there are plenty of directions that are worth pursuing. For instance, the dependency on the dimension d and the score estimation error remains sub-optimal, and more refined analyses are needed in order to tighten our result. Also, as mentioned previously, establishing sharp TV-type upper bound for



Images generated using vanilla DDPM.



Images generated using our method.

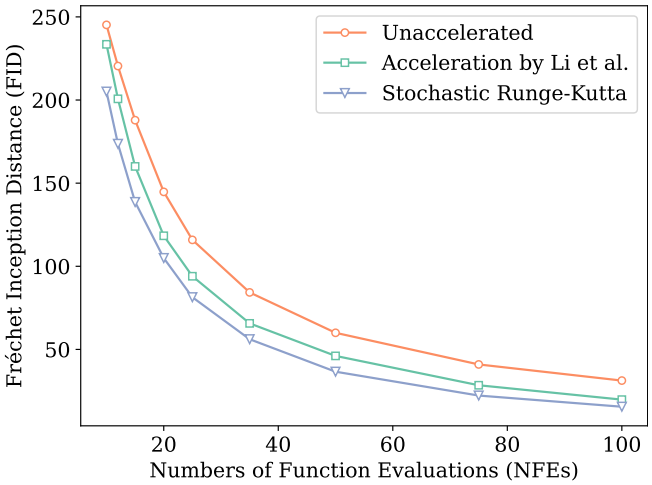


Figure 2: Simulation results using pre-trained score functions for the CIFAR-10 dataset. The left panel shows images generated by the vanilla DDPM and our method with 35 NFEs. The right panel plots the FID scores for all three methods across different NFEs.

Images generated using vanilla DDPM.

Images generated using our method.

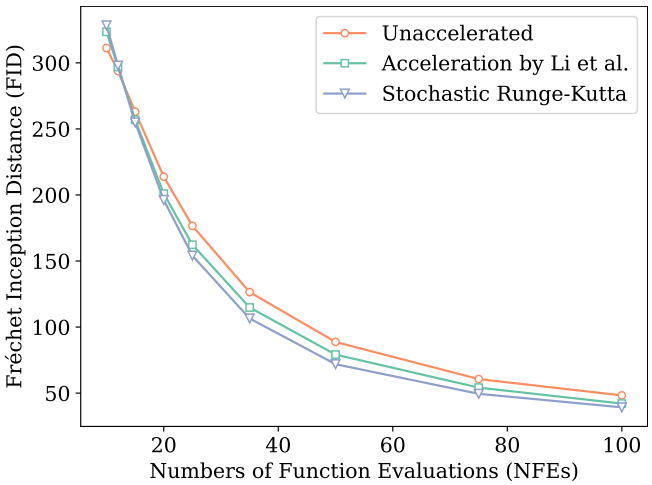


Figure 3: Simulation results using pre-trained score functions for the ImageNet-64 dataset. The left panel shows images generated by the vanilla DDPM and our method with 35 NFEs. The right panel plots the FID scores for all three methods across different NFEs.

our proposed sampler (as in Li and Yan (2024b) for DDPM) would be an interesting direction and call for new techniques, as the Girsanov-type arguments might not be applicable for analyzing the TV-distance. Furthermore, the recent work Li and Yan (2024a) has demonstrated the remarkable capability of DDPM in adapting to unknown low-dimensional structure; whether this appealing feature is inherited by our accelerated stochastic sampler is worth investigating. Finally, it would be important to develop fast and principled diffusion-based samplers that allow one to sample with guidance in a provably efficient manner (see, e.g., Wu et al. (2024); Chidambaram et al. (2024)).

Acknowledgements

Y. Chen is supported in part by the Alfred P. Sloan Research Fellowship, the AFOSR grant FA9550-22-1-0198, the ONR grant N00014-22-1-2354, and the NSF grant CCF-2221009. Y. Wei is supported in part by the NSF grants DMS-2147546/2015447, CAREER award DMS-2143215, CCF-2106778, CCF-2418156 and the Google Research Scholar Award. The authors gratefully acknowledge Timofey Efimov for his generous assistance with the numerical experiments.

A Technical lemmas

We collect in this section a couple of technical lemmas that are useful in establishing our main results.

Lemma A.1. *Denote by $(W_t)_{t\geq 0}$ a standard Brownian motion in $\mathbb{R}^d$. Then for all $t_k \leq t < t_{k+1}$, the covariance matrices of the following vectors are given by*

$$\begin{aligned}
&\mathrm{Cov}\left[\int_0^{t-t_k} e^{t-t_k-r}(W_{t_k+r}-W_{t_k})\mathrm{d}r\right] = \left[e^{2t-2t_k}/2 - 2e^{t-t_k} + (t-t_k+3/2)\right]\cdot I_d =: f_1(t-t_k)\cdot I_d,\\
&\mathrm{Cov}\left[\int_0^{t-t_k} e^{t-t_k-r}(W_{t_k+r}-W_{t_k})\mathrm{d}r,\, W_t - W_{t_k}\right] = \left[e^{t-t_k} - t + t_k - 1\right]\cdot I_d,\\
&\mathrm{Cov}\left[\int_0^{t-t_k} e^{t-t_k-r}\mathrm{d}W_{t_k+r}\right] = \left[e^{2(t-t_k)}/2 - 1/2\right]\cdot I_d =: f_2(t-t_k)\cdot I_d,\\
&\mathrm{Cov}\left[\int_0^{t-t_k} e^{t-t_k-r}(W_{t_k+r}-W_{t_k})\mathrm{d}r, \int_0^{t-t_k} e^{t-t_k-r}\mathrm{d}W_{t_k+r}\right] = \left[e^{2(t-t_k)}/2 - e^{t-t_k} + 1/2\right]\cdot I_d =: f_3(t-t_k)\cdot I_d.
\end{aligned}$$

Here, we define, for notational simplicity, the following functions: for any $\Delta > 0$, let

$$\begin{aligned}
f_1(\Delta) &= e^{2\Delta}/2 - 2e^{\Delta} + \Delta + 3/2,\\
f_2(\Delta) &= e^{2\Delta}/2 - 1/2,\\
f_3(\Delta) &= e^{2\Delta}/2 - e^{\Delta} + 1/2.
\end{aligned} \tag{32}$$

Proof of Lemma A.1. For $t_k \leq t < t_{k+1}$, set

$$\mathcal{H}(t) = \mathrm{Cov}\left[\int_0^{t-t_k} e^{t-t_k-r}(W_{t_k+r}-W_{t_k})\mathrm{d}r\right] = \mathbb{E}\left[\left(\int_0^{t-t_k} e^{t-t_k-r}(W_{t_k+r}-W_{t_k})\mathrm{d}r\right)^{\otimes 2}\right].$$

Observe that $\mathcal{H}(t_k) = 0_{d\times d}$. Taking the derivative of $\mathcal{H}(t)$ with respect to t, we reach

$$\begin{aligned}
\mathcal{H}'(t) &= \mathbb{E}\left[\int_0^{t-t_k} e^{t-t_k-r}(W_{t_k+r}-W_{t_k})\mathrm{d}r \otimes \left(W_t - W_{t_k} + \int_0^{t-t_k} e^{t-t_k-r}(W_{t_k+r}-W_{t_k})\mathrm{d}r\right)\right]\\
&\quad + \mathbb{E}\left[\left(W_t - W_{t_k} + \int_0^{t-t_k} e^{t-t_k-r}(W_{t_k+r}-W_{t_k})\mathrm{d}r\right) \otimes \int_0^{t-t_k} e^{t-t_k-r}(W_{t_k+r}-W_{t_k})\mathrm{d}r\right]\\
&= 2\mathcal{H}(t) + \left(2\int_0^{t-t_k} e^{t-t_k-r} r\,\mathrm{d}r\right) I_d
\end{aligned}$$

$$= 2\mathcal{H}(t) + 2(e^{t-t_k} - t + t_k - 1) \cdot I_d.$$

Solving the above ordinary differential equation yields

$$\mathcal{H}(t) = \Big(\frac{1}{2}e^{2t-2t_k} - 2e^{t-t_k} + (t - t_k + 3/2)\Big) I_d,$$

thereby completing the proof of the first advertised identity.

As for the second claimed identity, we make the observation that

$$\mathrm{Cov}\left[\int_0^{t-t_k} e^{t-t_k-r}(W_{t_k+r} - W_{t_k})\mathrm{d}r,\, W_t - W_{t_k}\right] = \left(\int_0^{t-t_k} e^{t-t_k-r} r \mathrm{d}r\right) I_d = \big(e^{t-t_k} - t + t_k - 1\big) \cdot I_d.$$

With regards to the third claimed identity, we have

$$\mathrm{Cov}\left[\int_0^{t-t_k} e^{t-t_k-r}\mathrm{d}W_{t_k+r}\right] = \left(\int_0^{t-t_k} e^{2(t-t_k-r)}\mathrm{d}r\right) I_d = \frac{1}{2}\big(e^{2(t-t_k)} - 1\big) I_d.$$

To prove the last result, it is seen that

$$\begin{aligned}\mathrm{Cov}\left[\int_0^{t-t_k} e^{t-t_k-r}(W_{t_k+r} - W_{t_k})\mathrm{d}r, \int_0^{t-t_k} e^{t-t_k-r}\mathrm{d}W_{t_k+r}\right] &= \left(\int_0^{t-t_k}\int_0^{t-t_k} e^{2(t-t_k)-r-s}\mathbb{1}_{s\le r}\mathrm{d}s\mathrm{d}r\right) I_d \\ &= \left(\int_0^{t-t_k}\int_0^{r} e^{2(t-t_k)-r-s}\mathrm{d}s\mathrm{d}r\right) I_d = \Big(\frac{1}{2}e^{2(t-t_k)} - e^{t-t_k} + \frac{1}{2}\Big) I_d.\end{aligned}$$

The proof is thus complete. □

Lemma A.2. *Consider any random object $M \in \mathbb{R}^{d\times d}$ and any random variable $\alpha \in \mathbb{R}$, as well as a filtration $\mathcal{F}$. Then, it holds that*

$$\mathbb{E}\left[\big\|\mathbb{E}[\alpha M \mid \mathcal{F}]\big\|_{\mathrm{F}}^4\right] \le \sqrt{\mathbb{E}[\alpha^8] \cdot \mathbb{E}\big[\|M\|_{\mathrm{F}}^8\big]}.$$

Proof of Lemma A.2. It follows from Cauchy–Schwarz that $\|\mathbb{E}[\alpha M \mid \mathcal{F}]\|_{\mathrm{F}}^2 \le \mathbb{E}[\alpha^2 \mid \mathcal{F}]\mathbb{E}[\|M\|_{\mathrm{F}}^2 \mid \mathcal{F}]$, which in turn yields

$$\begin{aligned}\mathbb{E}\left[\big\|\mathbb{E}[\alpha M \mid \mathcal{F}]\big\|_{\mathrm{F}}^4\right] &\le \mathbb{E}\left[\mathbb{E}[\alpha^2 \mid \mathcal{F}]^2\mathbb{E}\big[\|M\|_{\mathrm{F}}^2 \mid \mathcal{F}\big]^2\right] \le \mathbb{E}\left[\mathbb{E}[\alpha^2 \mid \mathcal{F}]^4\right]^{1/2}\mathbb{E}\left[\mathbb{E}[\|M\|_{\mathrm{F}}^2 \mid \mathcal{F}]^4\right]^{1/2} \\ &\le \sqrt{\mathbb{E}[\alpha^8]\mathbb{E}\big[\|M\|_{\mathrm{F}}^8\big]}.\end{aligned}$$

□

Lemma A.3. *For any $0 \le t_0 < t_0 + t < T$, the reverse process $(Y_t)_{0\le t\le T}$ (cf. Eq. (3)) obeys*

$$\begin{aligned}\mathbb{E}\big[\big\|Y_{t_0+t} - Y_{t_0} - \sqrt{2}B_{t_0+t} + \sqrt{2}B_{t_0}\big\|_2^2\big] &\lesssim d\sigma_{T-t_0-t}^{-2}t^2, \\ \mathbb{E}\big[\big\|Y_{t_0+t} - Y_{t_0} - \sqrt{2}B_{t_0+t} + \sqrt{2}B_{t_0}\big\|_2^4\big] &\lesssim d^2\sigma_{T-t_0-t}^{-4}t^4.\end{aligned}$$

Proof of Lemma A.3. For notational convenience, we define, for any $t \ge 0$, $\mathcal{D}(t) \coloneqq \mathbb{E}\big[\|Y_{t_0+t} - Y_{t_0} - \sqrt{2}B_{t_0+t} + \sqrt{2}B_{t_0}\|_2^2\big]$, which clearly obeys $\mathcal{D}(0) = 0$. It then follows that

$$\begin{aligned}|\mathcal{D}'(t)| &= 2\Big|\mathbb{E}\Big[\big\langle Y_{t_0+t} + 2s(t_0+t, Y_{t_0+t}), Y_{t_0+t} - Y_{t_0} - \sqrt{2}B_{t_0+t} + \sqrt{2}B_{t_0}\big\rangle\Big]\Big| \\ &\le 2\mathbb{E}\big[\big\|Y_{t_0+t} + 2s(t_0+t, Y_{t_0+t})\big\|_2^2\big]^{1/2}\mathcal{D}(t)^{1/2} \lesssim \sqrt{d}\,\sigma_{T-t_0-t}^{-1}\mathcal{D}(t)^{1/2}.\end{aligned}$$

Here, the first inequality arises from the Cauchy-Schwarz inequality, and the second inequality holds since

$$\mathbb{E}\big[\big\|Y_{t_0+t} + 2s(t_0+t, Y_{t_0+t})\big\|_2^2\big] \lesssim \mathbb{E}\big[\big\|Y_{t_0+t}\big\|_2^2\big] + \mathbb{E}\big[\big\|s(t_0+t, Y_{t_0+t})\big\|_2^2\big] = \mathbb{E}\big[\big\|X_{T-t_0-t}\big\|_2^2\big] + \mathbb{E}\big[\big\|s(t_0+t, X_{T-t_0-t})\big\|_2^2\big]$$

$$\lesssim d + d\sigma_{T-t_0-t}^{-2} \asymp d\sigma_{T-t_0-t}^{-2},$$

where the last line invokes Benton et al. (2024, Lemma 6) as well as Assumption A with $R = \sqrt{d}$. In view of the ODE comparison theorem, we see that $\mathcal{D}(t) \lesssim d\sigma_{T-t_0-t}^{-2} t^2$, thus establishing the first result of this lemma.

Similarly, we define $\overline{\mathcal{D}}(t) \coloneqq \mathbb{E}[\|Y_{t_0+t} - Y_{t_0} - \sqrt{2}B_{t_0+t} + \sqrt{2}B_{t_0}\|_2^4]$ with $\overline{\mathcal{D}}(0) = 0$. Taking the derivative of $\overline{\mathcal{D}}$, we see from Hölder's inequality that

$$\begin{aligned}
|\overline{\mathcal{D}}'(t)| &= 4\Big|\mathbb{E}\Big[\big\langle Y_{t_0+t} + 2s(t_0+t, Y_{t_0+t}), Y_{t_0+t} - Y_{t_0} - \sqrt{2}B_{t_0+t} + \sqrt{2}B_{t_0}\big\rangle \big\|Y_{t_0+t} - Y_{t_0} - \sqrt{2}B_{t_0+t} + \sqrt{2}B_{t_0}\big\|_2^2\Big]\Big| \\
&\le 4\mathbb{E}\big[\big\|Y_{t_0+t} + 2s(t_0+t, Y_{t_0+t})\big\|_2^4\big]^{1/4} \mathbb{E}\big[\big\|Y_{t_0+t} - Y_{t_0} - \sqrt{2}B_{t_0+t} + \sqrt{2}B_{t_0}\big\|_2^4\big]^{3/4} \\
&= 4\mathbb{E}\big[\big\|Y_{t_0+t} + 2s(t_0+t, Y_{t_0+t})\big\|_2^4\big]^{1/4} \overline{\mathcal{D}}(t)^{3/4} \\
&\lesssim \sqrt{d}\sigma_{T-t_0-t}^{-1} \overline{\mathcal{D}}(t)^{3/4},
\end{aligned}$$

which in turn implies that $\overline{\mathcal{D}}(t) \lesssim d^2\sigma_{T-t_0-t}^{-4} t^4$. This concludes the proof. □

B Properties of the score function

In this section, we gather several useful properties of the ground-truth score function $\{s(t,\cdot)\}$. Recall that the true score functions admit the following expression

$$s(T-t, x) = \frac{\lambda_t m(t,x) - x}{\sigma_t^2}, \qquad \text{with } m(t,x) = \mathop{\mathbb{E}}_{(\theta,g)\sim q_0 \otimes \mathcal{N}(0,I_d)} \big[\theta \mid \lambda_t\theta + \sigma_t g = x\big], \tag{33}$$

where we recall that $\lambda_t = e^{-t}$ and $\sigma_t = \sqrt{1-e^{-2t}}$. In addition, we find it convenient to define the function

$$f_0(\theta, x, t) \coloneqq \frac{\lambda_t^2}{\sigma_t^4}\|\theta\|_2^2 - \frac{\lambda_t + \lambda_t^3}{\sigma_t^4}\langle x, \theta\rangle. \tag{34}$$

Lemma B.1. *Recall that q_t is the distribution of $\lambda_t\theta + \sigma_t g$ for $(\theta, g) \sim q_0 \otimes \mathcal{N}(0, I_d)$. Under Assumption A,*

$$\partial_t \nabla_x \log q_t(x) = -\frac{\lambda_t(2-\sigma_t^2)}{\sigma_t^4} m(t,x) + \frac{2\lambda_t^2}{\sigma_t^4} x - \frac{\lambda_t^2(2-\sigma_t^2)}{\sigma_t^6} C_t(x)x + \frac{\lambda_t^3}{\sigma_t^6} v_t(x)$$

holds for any $x \in \mathbb{R}^d$ and $t \in (0, T]$, where

$$\begin{aligned}
m(t,x) &= \mathbb{E}[\theta \mid \lambda_t\theta + \sigma_t g = x] \in \mathbb{R}^d, \\
C_t(x) &= \mathrm{Cov}[\theta \mid \lambda_t\theta + \sigma_t g = x] \in \mathbb{R}^{d\times d}, \\
v_t(x) &= \mathbb{E}\big[\|\theta\|_2^2\big(\theta - m(t,x)\big) \mid \lambda_t\theta + \sigma_t g = x\big] \in \mathbb{R}^d.
\end{aligned}$$

Proof of Lemma B.1. Recall from Eq. (33) that

$$\nabla_x \log q_t(x) = s(T-t, x) = \frac{\lambda_t m(t,x) - x}{\sigma_t^2},$$

where $m(t,x) = \mathbb{E}[\theta \mid \lambda_t\theta + \sigma_t g = x]$ for $(\theta, g) \sim q_0 \otimes \mathcal{N}(0, I_d)$. To begin with, it is easily seen that

$$\partial_t(\sigma_t^{-2}) = -2\lambda_t^2\sigma_t^{-4}, \qquad \partial_t(\lambda_t\sigma_t^{-2}) = -\lambda_t(2-\sigma_t^2)\sigma_t^{-4}. \tag{35}$$

Next, we turn to $\partial_t m(t,x)$, and observe that

$$m(t,x) = \frac{\int \theta \exp(\lambda_t\sigma_t^{-2}\langle x,\theta\rangle - \lambda_t^2\sigma_t^{-2}\|\theta\|_2^2/2) q_0(\mathrm{d}\theta)}{\int \exp(\lambda_t\sigma_t^{-2}\langle x,\theta\rangle - \lambda_t^2\sigma_t^{-2}\|\theta\|_2^2/2) q_0(\mathrm{d}\theta)}.$$

Given our bounded fourth-moment assumption on q_0, we can readily apply Fubini's theorem and exchange the order of differentiation and integration, which leads to the following equation

$$\begin{aligned}\partial_t m(t,x) =& \frac{\lambda_t^2}{\sigma_t^4}\mathbb{E}\left[\|\theta\|_2^2\big(\theta - m(t,x)\big) \mid \lambda_t\theta + \sigma_t g = x\right] - \frac{\lambda_t(2-\sigma_t^2)}{\sigma_t^4}\mathbb{E}\big[\langle x,\theta\rangle\big(\theta - m(t,x)\big) \mid \lambda_t\theta + \sigma_t g = x\big] \\ =& \frac{\lambda_t^2}{\sigma_t^4} v_t(x) - \frac{\lambda_t(2-\sigma_t^2)}{\sigma_t^4} C_t(x)x. \end{aligned} \tag{36}$$

The proof can thus be completed by putting together Eqs. (35) and (36). □

Lemma B.2. *Suppose that the target distribution q_0 has finite fourth moment. Then, for all $t \in (0,T]$ and $x \in \mathbb{R}^d$, the following identities hold:*

1. $\nabla_x s(T-t,x) = -\sigma_t^{-2} I_d + \lambda_t^2\sigma_t^{-4}\operatorname{Cov}[\theta \mid \lambda_t\theta + \sigma_t g = x]$.
2. $\nabla_x^2 s(T-t,x) = \lambda_t^3\sigma_t^{-6}\mathbb{E}\big[\big(\theta - m(t,x)\big)\otimes\big(\theta - m(t,x)\big)\otimes\big(\theta - m(t,x)\big) \mid \lambda_t\theta + \sigma_t g = x\big]$.
3. $\partial_t s(T-t,x) = -(\lambda_t + \lambda_t^3)\sigma_t^{-4} m(t,x) + 2\lambda_t^2\sigma_t^{-4}x + \lambda_t\sigma_t^{-2}\mathbb{E}\big[(\theta - m(t,x))f_0(\theta,x,t) \mid \lambda_t\theta + \sigma_t g = x\big]$, *with $f_0(\theta,x,t)$ defined in Eq. (34). This expression can also be equivalently rewritten as*

$$\begin{aligned}\partial_t s(T-t,x) =& -\frac{\lambda_t}{\sigma_t^2}\mathbb{E}\big[\theta \mid \lambda_t\theta + \sigma_t g = x\big] + \frac{2\lambda_t^2}{\sigma_t^3}\mathbb{E}\big[g \mid \lambda_t\theta + \sigma_t g = x\big] \\ &+ \frac{\lambda_t}{\sigma_t^2}\mathbb{E}\big[\big(\theta - m(t,x)\big)f_0(\theta,x,t) \mid \lambda_t\theta + \sigma_t g = x\big].\end{aligned}$$

Proof of Lemma B.2. The first identity has been established in, e.g., Benton et al. (2024, Lemma 5).

To establish the second identity claimed in the lemma, we observe that: from our assumption that q_0 has bounded fourth moment, we can apply Fubini's theorem and exchange the order of differentiation and integration to obtain

$$\begin{aligned}\nabla_x^2 s(T-t,x) &= \lambda_t^2\sigma_t^{-4}\nabla_x \operatorname{Cov}\left[\theta \mid \lambda_t\theta + \sigma_t g = x\right] \\ &= \lambda_t^3\sigma_t^{-6}\mathbb{E}\big[\big(\theta - m(t,x)\big)\otimes\big(\theta - m(t,x)\big)\otimes\big(\theta - m(t,x)\big) \mid \lambda_t\theta + \sigma_t g = x\big].\end{aligned}$$

This proves the second point of the lemma.

Finally, we prove the third identity claimed in the lemma. Invoking Lemma B.1 gives

$$\begin{aligned}\partial_t s(T-t,x) = \partial_t\nabla_x \log q_t(x) &= -\frac{\lambda_t(2-\sigma_t^2)}{\sigma_t^4}m(t,x) + \frac{2\lambda_t^2}{\sigma_t^4}x - \frac{\lambda_t^2(2-\sigma_t^2)}{\sigma_t^6}C_t(x)x + \frac{\lambda_t^3}{\sigma_t^6}v_t(x) \\ &= -\frac{\lambda_t+\lambda_t^3}{\sigma_t^4}m(t,x) + \frac{2\lambda_t^2}{\sigma_t^4}x + \frac{\lambda_t}{\sigma_t^2}\mathbb{E}\big[\big(\theta - m(t,x)\big)f_0(\theta,x,t) \mid \lambda_t\theta + \sigma_t g = x\big],\end{aligned}$$

where $C_t(x) \coloneqq \operatorname{Cov}[\theta \mid \lambda_t\theta + \sigma_t g = x]$ and $v_t(x) \coloneqq \mathbb{E}\big[\|\theta\|_2^2\big(\theta - m(t,x)\big) \mid \lambda_t\theta + \sigma_t g = x\big]$ with $(\theta, g) \sim q_0 \otimes \mathcal{N}(0, I_d)$. The proof is complete by taking advantage of the identity $x = \lambda_t m(t,x) + \sigma_t\mathbb{E}[g \mid \lambda_t\theta + \sigma_t g = x]$. □

Lemma B.3. *Assume that the target distribution q_0 has finite fourth moment. Then, for all $t \in (0,T]$ and $x \in \mathbb{R}^d$, it holds that*

$$\begin{aligned}\partial_t\nabla_x s(T-t,x) =& \frac{2\lambda_t^2}{\sigma_t^4}I_d + \left(\frac{2\lambda_t^2}{\sigma_t^4} - \frac{4\lambda_t^2}{\sigma_t^6}\right)\operatorname{Cov}[\theta \mid \lambda_t\theta + \sigma_t g = x] \\ &+ \frac{\lambda_t^2}{\sigma_t^4}\mathbb{E}\left[\big(\theta - m(t,x)\big)\big(\theta - m(t,x)\big)^\top\big(f_0(\theta,x,t) - \mathbb{E}\big[f_0(\theta,x,t) \mid \lambda_t\theta + \sigma_t g = x\big]\big) \mid \lambda_t\theta + \sigma_t g = x\right].\end{aligned}$$

Proof of Lemma B.3. Since q_0 has bounded fourth moment, by Fubini's theorem we can exchange the order of differentiation and integration. In addition, by the third point of Lemma B.2, we know that

$$\partial_t s(T-t,x) = -(\lambda_t + \lambda_t^3)\sigma_t^{-4}m(t,x) + 2\lambda_t^2\sigma_t^{-4}x + \lambda_t\sigma_t^{-2}\mathbb{E}\big[\big(\theta - m(t,x)\big)f_0(\theta,x,t) \mid \lambda_t\theta + \sigma_t g = x\big].$$

As a consequence, we can compute

$$\nabla_x \partial_t s(T-t,x) = \frac{2\lambda_t^2}{\sigma_t^4} I_d - \frac{2\lambda_t^2 + 2\lambda_t^4}{\sigma_t^6} \operatorname{Cov}[\theta \mid \lambda_t \theta + \sigma_t g = x]$$
$$+ \frac{\lambda_t^2}{\sigma_t^4} \mathbb{E}\big[\big(\theta - m(t,x)\big)\big(\theta - m(t,x)\big)^\top \big(f_0(\theta,x,t) - \mathbb{E}[f_0(\theta,x,t) \mid \lambda_t\theta + \sigma_t g = x]\big) \mid \lambda_t \theta + \sigma_t g = x\big]$$

as claimed. □

Lemma B.4. *Assume that the target distribution q_0 has finite fourth moment. Then, for all $t \in (0,T]$ and $x \in \mathbb{R}^d$, it holds that*

$$\partial_t \nabla_x^2 s(T-t,x) = -\frac{3(\lambda_t^3 + \lambda_t^5)}{\sigma_t^8} \mathbb{E}\big[\big(\theta - m(t,x)\big)^{\otimes 3} \mid \lambda_t\theta + \sigma_t g = x\big] - \frac{\lambda_t^3}{2\sigma_t^6} \sum_{\pi \in \mathrm{perm}(3)} M_\pi$$
$$+ \frac{\lambda_t^3}{\sigma_t^6} \mathbb{E}\Big[\big(\theta - m(t,x)\big)^{\otimes 3} \big(f_0(\theta,x,t) - \mathbb{E}[f_0(\theta,x,t) \mid \lambda_t\theta + \sigma_t g = x]\big) \mid \lambda_t\theta + \sigma_t g = x\Big],$$

where for $\pi \in \mathrm{perm}(3)$ and $i_1, i_2, i_3 \in [d]$, we take

$$(M_\pi)_{i_1 i_2 i_3} = \mathbb{E}[(\theta_{i_{\pi(1)}} - m_{i_{\pi(1)}})(\theta_{i_{\pi(2)}} - m_{i_{\pi(2)}}) \mid \lambda_t\theta + \sigma_t g = x]$$
$$\times \mathbb{E}[(\theta_{i_{\pi(3)}} - m_{i_{\pi(3)}})(f_0(\theta,x,t) - m_{f_0}) \mid \lambda_t\theta + \sigma_t g = x].$$

In the above display,

$$m_i = \mathbb{E}[\theta_i \mid \lambda_t\theta + \sigma_t g = x] \qquad \text{and} \qquad m_{f_0} = \mathbb{E}[f_0(\theta,x,t) \mid \lambda_t\theta + \sigma_t g = x].$$

Proof of Lemma B.4. With the bounded fourth-moment assumption on q_0 in place, one can apply Fubini's theorem to swap the order of differentiation and integration and obtain

$$\partial_t \nabla_x^2 s(T-t,x) = \nabla_x \partial_t \nabla_x s(T-t,x)$$
$$= -\frac{2\lambda_t^2 + 2\lambda_t^4}{\sigma_t^6} \nabla_x \operatorname{Cov}\big[\theta \mid \lambda_t\theta + \sigma_t g = x\big]$$
$$+ \frac{\lambda_t^2}{\sigma_t^4} \nabla_x \mathbb{E}\Big[\big(\theta - m(t,x)\big)\big(\theta - m(t,x)\big)^\top \big(f_0(\theta,x,t) - \mathbb{E}\big[f_0(\theta,x,t) \mid \lambda_t\theta + \sigma_t g = x\big]\big) \mid \lambda_t\theta + \sigma_t g = x\Big]$$
$$= -\frac{3(\lambda_t^3 + \lambda_t^5)}{\sigma_t^8} \mathbb{E}\big[\big(\theta - m(t,x)\big)^{\otimes 3} \mid \lambda_t\theta + \sigma_t g = x\big] - \frac{\lambda_t^3}{2\sigma_t^6} \sum_{\pi \in \mathrm{perm}(3)} M_\pi$$
$$+ \frac{\lambda_t^3}{\sigma_t^6} \mathbb{E}\big[(\theta - m(t,x))^{\otimes 3} (f_0(\theta,x,t) - \mathbb{E}[f_0(\theta,x,t) \mid \lambda_t\theta + \sigma_t g = x]) \mid \lambda_t\theta + \sigma_t g = x\big]$$

as claimed. □

Lemma B.5. *Assume that the target distribution q_0 has finite fourth moment. Then, for all $t \in (0,T]$ and $x \in \mathbb{R}^d$, it holds that*

$$\nabla_x^3 s(T-t,x) = \frac{\lambda_t^4}{\sigma_t^8} \Big(\mathbb{E}\big[\big(\theta - m(t,x)\big)^{\otimes 4} \mid \lambda_t\theta + \sigma_t g = x\big] - \mathcal{T}\big(\operatorname{Cov}[\theta \mid \lambda_t\theta + \sigma_t g = x]^{\otimes 2}\big)\Big) \in \mathbb{R}^{d^4}.$$

Here, for any tensor $X \in \mathbb{R}^{d\times d\times d\times d}$, we take $\mathcal{T}(X) = \sum_{\pi\in\mathrm{perm}(4)} \mathcal{T}_\pi(X)/8$ and $\mathcal{T}_\pi(X) \in \mathbb{R}^{d\times d\times d\times d}$, such that $\mathcal{T}_\pi(X)_{i_1 i_2 i_3 i_4} = X_{i_{\pi(1)} i_{\pi(2)} i_{\pi(3)} i_{\pi(4)}}$.

Proof of Lemma B.5. Let us invoke Fubini's theorem to swap the order of differentiation and integration, thus leading to

$$\nabla_x^3 s(T-t,x) = \lambda_t^3 \sigma_t^{-6} \nabla_x \mathbb{E}[(\theta - m(t,x)) \otimes (\theta - m(t,x)) \otimes (\theta - m(t,x)) \mid \lambda_t\theta + \sigma_t g = x]$$

$$= \lambda_t^4 \sigma_t^{-8} \cdot \Big(\mathbb{E}[(\theta - m(t,x))^{\otimes 4} \mid \lambda_t \theta + \sigma_t g = x] - \mathcal{T}(\mathrm{Cov}[\theta \mid \lambda_t \theta + \sigma_t g = x]^{\otimes 2}) \Big)$$

The proof is thus complete. □

Lemma B.6. *We assume Assumption A. Recall that X_{T-t} is defined in Eq. (3) and has marginal distribution q_t. Then, for $t \in (0, T]$, it holds that*

$$\mathbb{E}\left[\left\| \partial_t \nabla_x \log q_t(X_{T-t}) \right\|_2^2 \right] \lesssim \frac{d^3}{\sigma_t^6}.$$

Proof of Lemma B.6. Invoking Lemma B.1, we obtain the following upper bound:

$$\begin{aligned} &\mathbb{E}\left[\left\| \partial_t \nabla_x \log q_t(X_{T-t}) \right\|_2^2 \right] \\ &\lesssim \underbrace{\mathbb{E}\left[\left\| -\frac{\lambda_t(2-\sigma_t^2)}{\sigma_t^4} m(t, X_{T-t}) + \frac{2\lambda_t^2}{\sigma_t^4} X_{T-t} \right\|_2^2 \right]}_{(i)} + \underbrace{\mathbb{E}\left[\left\| -\frac{\lambda_t^2(2-\sigma_t^2)}{\sigma_t^6} C_t(X_{T-t}) X_{T-t} + \frac{\lambda_t^3}{\sigma_t^6} v_t(X_{T-t}) \right\|_2^2 \right]}_{(ii)}. \end{aligned} \tag{37}$$

We shall bound terms (i) and (ii) in Eq. (37) separately in the sequel.

Let us start with term (i), for which we observe that

$$-\frac{\lambda_t(2-\sigma_t^2)}{\sigma_t^4} m(t, X_{T-t}) + \frac{2\lambda_t^2}{\sigma_t^4} X_{T-t} = -\frac{\lambda_t}{\sigma_t^2} m(t, X_{T-t}) + \frac{2\lambda_t^2}{\sigma_t^3} \mathbb{E}\big[g \mid \lambda_t \theta + \sigma_t g = X_{T-t} \big]. \tag{38}$$

By Jensen's inequality, we have

$$\mathbb{E}\left[\|m(t, X_{T-t})\|_2^2 \right] \le \mathbb{E}\left[\|\theta\|_2^2 \right] \le d, \qquad \mathbb{E}\big[\|\mathbb{E}[g \mid \lambda_t \theta + \sigma_t g = X_{T-t}]\|_2^2 \big] \le \mathbb{E}\left[\|g\|_2^2 \right] = d, \tag{39}$$

where $(\theta, g) \sim q_0 \otimes \mathcal{N}(0, I_d)$. Substituting Eqs. (38) and (39) into term (i), we arrive at

$$(i) \lesssim \frac{\lambda_t^2 d}{\sigma_t^4} + \frac{\lambda_t^4 d}{\sigma_t^6}. \tag{40}$$

Next, we turn attention to term (ii). To this end, write $X_{T-t} = \lambda_t \Theta + \sigma_t G$, where $(\Theta, G) \sim q_0 \otimes \mathcal{N}(0, I_d)$. Note that

$$\begin{aligned} &-\frac{\lambda_t^2(2-\sigma_t^2)}{\sigma_t^6} C_t(X_{T-t}) X_{T-t} + \frac{\lambda_t^3}{\sigma_t^6} v_t(X_{T-t}) \\ &= -\frac{\lambda_t^2 + \lambda_t^4}{\sigma_t^5} C_t(X_{T-t}) G + \frac{\lambda_t^3}{\sigma_t^6} \mathbb{E}\left[\big(\theta - m(t, X_{T-t})\big) \big(\theta - m(t, X_{T-t})\big)^\top \theta \mid \lambda_t \theta + \sigma_t g = X_{T-t} \right] \\ &\quad - \frac{\lambda_t^3}{\sigma_t^6} C_t(X_{T-t}) \big(\Theta - m(t, X_{T-t})\big) - \frac{\lambda_t^5}{\sigma_t^6} C_t(X_{T-t}) \Theta \\ &= -\frac{\lambda_t^2 + \lambda_t^4}{\sigma_t^5} C_t(X_{T-t}) G + \frac{\lambda_t^3}{\sigma_t^6} \mathbb{E}\left[\big(\theta - m(t, X_{T-t})\big) \big\|\theta - m(t, X_{T-t})\big\|_2^2 \mid \lambda_t \theta + \sigma_t g = X_{T-t} \right] \\ &\quad - \frac{\lambda_t^3 + \lambda_t^5}{\sigma_t^6} C_t(X_{T-t}) \big(\Theta - m(t, X_{T-t})\big) + \frac{\lambda_t^3}{\sigma_t^4} C_t(X_{T-t}) m(t, X_{T-t}). \end{aligned} \tag{41}$$

We then separately upper bound the terms in the lase line of Eq. (41). To this end, we find the following expressions useful (recall that $X_{T-t} = \lambda_t \Theta + \sigma_t G$ for $(\Theta, G) \sim q_0 \otimes \mathcal{N}(0, 1)$):

$$\begin{aligned} &C_t(X_{T-t}) = \frac{\sigma_t^2}{\lambda_t^2} \mathbb{E}\left[\big(g - \mathbb{E}[g \mid \lambda_t \theta + \sigma_t g = X_{T-t}]\big) \big(g - \mathbb{E}[g \mid \lambda_t \theta + \sigma_t g = X_{T-t}]\big)^\top \mid \lambda_t \theta + \sigma_t g = X_{T-t} \right], \\ &\mathbb{E}\left[\big(\theta - m(t, X_{T-t})\big) \big\|\theta - m(t, X_{T-t})\big\|_2^2 \mid \lambda_t \theta + \sigma_t g = X_{T-t} \right] \\ &\qquad = -\frac{\sigma_t^3}{\lambda_t^3} \mathbb{E}\left[\big(g - \mathbb{E}[g \mid \lambda_t \theta + \sigma_t g = X_{T-t}]\big) \big\|g - \mathbb{E}[g \mid \lambda_t \theta + \sigma_t g = X_{T-t}]\big\|_2^2 \mid \lambda_t \theta + \sigma_t g = X_{T-t} \right], \\ &\Theta - m_t(X_{T-t}) = -\frac{\sigma_t}{\lambda_t} \big(G - \mathbb{E}[g \mid \lambda_t \theta + \sigma_t g = X_{T-t}]\big). \end{aligned} \tag{42}$$

- Let us look at the first summand in the last line of Eq. (41), namely, the term $-\sigma_t^{-5}(\lambda_t^2+\lambda_t^4)C_t(X_{T-t})G$. In view of Eq. (42), we can deduce that

$$\mathbb{E}\left[\left\|C_t(X_{T-t})G\right\|_2^2\right] \overset{(a)}{\leq} \mathbb{E}\left[\left\|C_t(X_{T-t})\right\|_{\mathrm{F}}^4\right]^{1/2}\mathbb{E}\left[\left\|G\right\|_2^4\right]^{1/2} \overset{(b)}{\lesssim} \frac{\sigma_t^4 d^3}{\lambda_t^4}, \tag{43}$$

  where step (a) is by the Cauchy–Schwarz inequality, and step (b) arises from the Jensen inequality. Using Eq. (43), we see that

$$\frac{(\lambda_t^2+\lambda_t^4)^2}{\sigma_t^{10}}\mathbb{E}\left[\left\|C_t(X_{T-t})G\right\|_2^2\right] \lesssim \frac{d^3}{\sigma_t^6}. \tag{44}$$

- Regarding the second summand in Eq. (41), note that $\lambda_t^3\sigma_t^{-6}\mathbb{E}[(\theta-m(t,X_{T-t}))\|\theta-m(t,X_{T-t})\|_2^2 \mid \lambda_t\theta+\sigma_t g=X_{T-t}]$. Applying Eq. (42), the Cauchy–Schwarz inequality and Jensen's inequality yields

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbb{E}\left[(\theta-m(t,X_{T-t}))\|\theta-m(t,X_{T-t})\|_2^2 \mid \lambda_t\theta+\sigma_t g=X_{T-t}\right]\right\|_2^2\right] \\
&\leq \mathbb{E}\left[\mathbb{E}[\|\theta-m(t,X_{T-t})\|_2^2 \mid \lambda_t\theta+\sigma_t g=X_{T-t}]\cdot\mathbb{E}[\|\theta-m(t,X_{T-t})\|_2^4 \mid \lambda_t\theta+\sigma_t g=X_{T-t}]\right] \\
&\leq \mathbb{E}\left[\|\Theta-m(t,X_{T-t})\|_2^4\right]^{1/2}\cdot\mathbb{E}\left[\|\Theta-m(t,X_{T-t})\|_2^8\right]^{1/2} \lesssim \frac{\sigma_t^6 d^3}{\lambda_t^6},
\end{aligned}$$

  which further implies that

$$\frac{\lambda_t^6}{\sigma_t^{12}}\mathbb{E}\left[\|\mathbb{E}[(\theta-m(t,X_{T-t}))\|\theta-m(t,X_{T-t})\|_2^2 \mid \lambda_t\theta+\sigma_t g=X_{T-t}]\|_2^2\right] \lesssim \frac{d^3}{\sigma_t^6}. \tag{45}$$

- The remaining two terms in Eq. (41) can be controlled in a similar manner. The proof idea is similar to that for the first two terms, and we skip a detailed explanation for the compactness of presentation. Specifically, we obtain the following upper bounds:

$$\frac{(\lambda_t^3+\lambda_t^5)^2}{\sigma_t^{12}}\mathbb{E}\left[\|C_t(X_{T-t})(\Theta-m(t,X_{T-t}))\|_2^2\right] \lesssim \frac{d^3}{\sigma_t^6}, \tag{46}$$

$$\frac{\lambda_t^6}{\sigma_t^8}\mathbb{E}\left[\|C_t(X_{T-t})m(t,X_{T-t})\|_2^2\right] \lesssim \frac{\lambda_t^2 d^3}{\sigma_t^4}. \tag{47}$$

Finally, we put together Eqs. (44) to (47). Invoking Eq. (41), we can then conclude that

$$(ii) \lesssim \frac{d^3}{\sigma_t^6}. \tag{48}$$

To finish up, combine Eqs. (40) and (48) to obtain

$$\mathbb{E}\left[\left\|\partial_t\nabla_x \log q_t(X_{T-t})\right\|_2^2\right] \lesssim \frac{d^3}{\sigma_t^6},$$

thus concluding the proof. □

C Bounding the KL divergence between diffusion processes

C.1 Proof of Lemma 4.1

First, we show that for any $\tau\in(0,t_{k+1}-t_k)$, there exists a unique strong solution to the SDE (29) on the interval $[t_k+\tau,t_{k+1}]$. To this end, we shall introduce an augmented process, and show that $(H_s^\tau)_{t_k\leq s\leq t}$ is a subset of this process. Let us begin by determining the drift function of this process and proving its Lipschitz

continuity. Recall that $\overline{\mathcal{F}}$ is defined as the drift functional of process (25a). For all $t \in (t_k, t_{k+1}]$, observe that we can write

$$\overline{\mathcal{F}}\big(t, Y_{t_k}, (H_s^\tau)_{t_k \le s \le t}\big) = \mathcal{G}_t\left(Y_{t_k}, H_t^\tau, \int_0^{t-t_k} e^{t-t_k-r} \mathrm{d}H_{t_k+r}^\tau, \int_0^{t-t_k} e^{t-t_k-r} H_{t_k+r}^\tau \mathrm{d}r\right).$$

for some continuous mapping $\mathcal{G}_t : \mathbb{R}^{4d} \to \mathbb{R}^d$. By the first point of Lemma B.2, we see that under Assumption A, for all $t \in [0, T)$ the mapping $x \mapsto s(t, x)$ is Lipschitz continuous. As a consequence, for any $\tau \in (0, t_{k+1} - t_k)$ and all $t \in [t_k + \tau, t_{k+1}]$, $\mathcal{G}_t$ is C_τ-Lipschitz continuous for some $C_\tau \in (0, \infty)$ that depends only on τ.

We then introduce an augmented process $L_t = (L_{1,t}, L_{2,t}, L_{3,t}, L_{4,t}) \in \mathbb{R}^{4d}$, defined as the solution to the following SDE:

$$\begin{aligned} \mathrm{d}L_t &= \begin{bmatrix} L_{1,t} + 2s(t, L_{1,t}) \\ \big(L_{1,t} + 2s(t, L_{1,t}) - \mathcal{G}_t(L_{1,t}, L_{2,t}, L_{3,t}, L_{4,t})\big)/\sqrt{2} \\ L_{3,t} + \big(L_{1,t} + 2s(t, L_{1,t}) - \mathcal{G}_t(L_{1,t}, L_{2,t}, L_{3,t}, L_{4,t})\big)/\sqrt{2} \\ L_{2,t} + L_{4,t} \end{bmatrix} \mathrm{d}t + \begin{bmatrix} \sqrt{2}\mathrm{d}B_t \\ \mathrm{d}B_t \\ \mathrm{d}B_t \\ 0_d \end{bmatrix} \\ &= b(t, L_t)\mathrm{d}t + \sigma(t, L_t)\mathrm{d}B_t. \end{aligned} \tag{49}$$

Here, $b(t, L) \in \mathbb{R}^{4d}$ and $\sigma(t, L) \in \mathbb{R}^{4d \times d}$. Since $\mathcal{G}_t$ is C_τ-Lipschitz continuous for all $t \in [t_k + \tau, t_{k+1}]$, we obtain that the mappings $L \mapsto b(t, L)$ and $L \mapsto \sigma(t, L)$ are Lipschitz continuous with a uniformly upper bounded Lipschitz constant for all $t \in [t_k + \tau, t_{k+1}]$. By Le Gall (2016, Theorem 8.3), SDE (49) has a unique strong solution, regardless of the initialization. This establishes the existence and uniqueness of process $(H_t^\tau)_{t_k \le t \le t_{k+1}}$ as a solution to (29).

In what follows, denote by $Q_k(y)$ the law of $(Y_t)_{t_k \le t \le t_{k+1}}$, and $\overline{Q}_k(y)$ the law of $(\mathcal{A}_t(Y_{t_k}))_{t_k \le t \le t_{k+1}}$, conditioned on $Y_{t_k} = y$, where

$$\mathcal{A}_t(y) := e^{t-t_k} y + (e^{t-t_k} - e^{-t+t_k}) s(t_k, y + g_{t_k,t}) + \sqrt{2} \int_0^{t-t_k} e^{t-t_k-r} \mathrm{d}W_{t_k+r}.$$

In the above display, we recall that $(W_t)_{t \ge 0}$ represents a d-dimensional standard Brownian motion, and $g_{t_k,t}$ is defined in Eq. (10b). Using the decomposition of KL divergence, we obtain

$$\mathsf{KL}\big(Q_{T-\delta} \,\|\, \overline{Q}_{T-\delta}\big) = \sum_{k=0}^{K-1} \mathbb{E}_{Q_{T-\delta}}\Big[\mathsf{KL}\big(Q_k(Y_{t_k}) \,\|\, \overline{Q}_k(Y_{t_k})\big)\Big],$$

where the expectation is taken with respect to $(Y_t)_{0 \le t \le T-\delta} \sim Q_{T-\delta}$.

For $\tau \in (0, t_{k+1} - t_k)$, we define the process $(U_t^\tau)_{t_k \le t \le t_{k+1}}$ with $U_{t_k}^\tau = Y_{t_k}$, and

$$\begin{aligned} \mathrm{d}U_t^\tau &= \big(U_t^\tau + 2s(t, U_t^\tau)\big)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, && \text{for } t_k \le t \le t_k + \tau, \\ \mathrm{d}U_t^\tau &= \overline{\mathcal{F}}\big(t, Y_{t_k}, (W_s - W_{t_k})_{t_k \le s \le t}\big)\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, && \text{for } t_k + \tau \le t \le t_{k+1}. \end{aligned}$$

As $\tau \to 0^+$, it holds that $\sup_{t_k \le t \le t_{k+1}} \|U_t^\tau - \mathcal{A}_t(Y_{t_k})\|_2 \overset{\text{a.s.}}{\to} 0$. Denote by $\overline{Q}_k^\tau(Y_{t_k})$ the conditional distribution of $(U_t^\tau)_{t_k \le t \le t_{k+1}}$ given $U_{t_k}^\tau = Y_{t_k}$. Therefore, for all $Y_{t_k} \in \mathbb{R}^d$ the distribution $\overline{Q}_k^\tau(Y_{t_k})$ converges weakly to $\overline{Q}_k(Y_{t_k})$. Invoking the same Girsanov-type arguments as in Chen et al. (2023c, Section 5.2), we see that

$$\mathsf{KL}\big(Q_k(Y_{t_k}) \,\|\, \overline{Q}_k^\tau(Y_{t_k})\big) = \int_{t_k+\tau}^{t_{k+1}} \mathbb{E}\Big[\big\|\mathcal{F}(t, Y_t) - \overline{\mathcal{F}}(t, Y_{t_k}, (H_s^\tau)_{t_k \le s \le t})\big\|_2^2 \,\big|\, Y_{t_k}\Big]\mathrm{d}t. \tag{50}$$

Leveraging the lower semicontinuity of KL divergence, we obtain

$$\begin{aligned} \mathbb{E}_{Q_{T-\delta}}\Big[\mathsf{KL}\big(Q_k(Y_{t_k}) \,\|\, \overline{Q}_k(Y_{t_k})\big)\Big] &\le \mathbb{E}_{Q_{T-\delta}}\Big[\liminf_{\tau \to 0^+} \mathsf{KL}\big(Q_k(Y_{t_k}) \,\|\, \overline{Q}_k^\tau(Y_{t_k})\big)\Big] \\ &\le \liminf_{\tau \to 0^+} \mathbb{E}_{Q_{T-\delta}}\Big[\mathsf{KL}\big(Q_k(Y_{t_k}) \,\|\, \overline{Q}_k^\tau(Y_{t_k})\big)\Big] \end{aligned}$$

$$= \liminf_{\tau \to 0^+} \int_{t_k+\tau}^{t_{k+1}} \mathbb{E}\Big[\big\| \mathcal{F}(t, Y_t) - \overline{\mathcal{F}}(t, Y_{t_k}, (H_s^\tau)_{t_k \le s \le t}) \big\|_2^2 \,\big|\, Y_{t_k} \Big] \mathrm{d}t,$$

where the second inequality above follows from Fatou's Lemma, and the last equality arises from Eq. (50). This completes the proof.

C.2 Proof of Lemma 4.2

By virtue of Eq. (29), for all $t \in [t_k + \tau, t_{k+1}]$ we have

$$H_t^\tau - H_{t_k}^\tau = B_t - B_{t_k} + v_{a,t} + v_{b,t} + v_{c,t} + v_{d,t} + v_{e,t} + v_{f,t}, \tag{51}$$

where the residual terms $v_{a,t}, v_{b,t}, v_{c,t}, v_{d,t}, v_{e,t}, v_{f,t} \in \mathbb{R}^d$ are defined respectively as follows:

$$\begin{aligned}
v_{a,t} &= \frac{1}{\sqrt{2}} \int_{t_k+\tau}^{t} (e^{\zeta - t_k} - e^{-\zeta + t_k}) s(t_k, Y_{t_k}) \mathrm{d}\zeta, \\
v_{b,t} &= \int_{t_k+\tau}^{t} \int_0^{\zeta - t_k} e^{\zeta - t_k - r} (\mathrm{d}B_{t_k+r} - \mathrm{d}H_{t_k+r}^\tau) \mathrm{d}\zeta, \\
v_{c,t} &= -\frac{1}{\sqrt{2}} \int_{t_k+\tau}^{t} (e^{\zeta - t_k} + e^{-\zeta + t_k}) \big(s(t_k, Y_{t_k} + h_{t_k,\zeta}) - s(t_k, Y_{t_k}) - \nabla_x s(t_k, Y_{t_k} + h_{t_k,\zeta}) h_{t_k,\zeta} \big) \mathrm{d}\zeta, \\
v_{d,t} &= \sqrt{2} \int_{t_k+\tau}^{t} \int_0^{\zeta - t_k} e^{\zeta - t_k - r} \big(s(t_k + r, Y_{t_k+r}) - s(t_k, Y_{t_k}) \big) \mathrm{d}r \mathrm{d}\zeta, \\
v_{e,t} &= -2 \int_{t_k+\tau}^{t} \nabla_x s(t_k, Y_{t_k} + h_{t_k,\zeta}) \int_0^{\zeta - t_k} e^{\zeta - t_k - r} H_{t_k+r}^\tau \mathrm{d}r \mathrm{d}\zeta, \\
v_{f,t} &= \sqrt{2} \int_{t_k+\tau}^{t} \big(s(\zeta, Y_\zeta) - s(t_k, Y_{t_k}) - \sqrt{2} \nabla_x s(t_k, Y_{t_k} + h_{t_k,\zeta}) H_\zeta^\tau \big) \mathrm{d}\zeta,
\end{aligned}$$

In the above display, we let

$$h_{t_k,\zeta} = \frac{2\sqrt{2}}{e^{\zeta - t_k} - e^{-\zeta + t_k}} \int_0^{\zeta - t_k} e^{\zeta - t_k - r} H_{t_k+r}^\tau \mathrm{d}r,$$

which essentially replaces $W_{t_k+r} - W_{t_k}$ with $H_{t_k+r}^\tau$ in the definition of $g_{t_k,\zeta}$.

Before proceeding to bounding the terms in Eq. (51), we find it helpful to first make some observations. For $t \in [t_k, t_{k+1}]$, define $\gamma_{t_k,t} = \|H_t^\tau - B_t + B_{t_k}\|_2$, which clearly obeys $\gamma_{t_k,t_k} = 0$. In view of the first point of Lemma B.2 and Assumption A, we see that

$$\big\| \nabla_x s(t,x) \big\|_2 \le 4 \big(\sigma_{T-t}^{-2} + \lambda_{T-t}^2 \sigma_{T-t}^{-4} \big) d$$

for all $x \in \mathbb{R}^d$ and $t \in [0, T)$. Similarly, by the second point of Lemma B.2, we can deduce that

$$\big\| \nabla_x^2 s(t,x) \big\|_2 \le 8 \lambda_{T-t}^3 \sigma_{T-t}^{-6} d^{3/2}$$

for all $x \in \mathbb{R}^d$ and $t \in [0, T)$.

Next, we adopt these upper bounds to analyze the terms on the right-hand side of Eq. (51). We first look at $v_{c,t}$. By the fundamental theorem of calculus, we have

$$s(t_k, Y_{t_k} + h_{t_k,\zeta}) - s(t_k, Y_{t_k}) = \int_0^1 \nabla_x s(t_k, Y_{t_k} + \eta h_{t_k,\zeta}) h_{t_k,\zeta} \mathrm{d}\eta.$$

Using this decomposition and the triangle inequality, we obtain that

$$\big\| s(t_k, Y_{t_k} + h_{t_k,\zeta}) - s(t_k, Y_{t_k}) - \nabla_x s(t_k, Y_{t_k} + h_{t_k,\zeta}) h_{t_k,\zeta} \big\|_2$$

$$\leq \int_0^1 \|\nabla_x s(t_k, Y_{t_k} + \eta h_{t_k,\zeta})\|_2 \|h_{t_k,\zeta}\|_2 \mathrm{d}\eta$$
$$\leq 8(\sigma_{T-t_k}^{-2} + \lambda_{T-t_k}^2 \sigma_{T-t_k}^{-4}) d \|h_{t_k,\zeta}\|_2.$$

Recall that under Assumption B, we have $\Delta_k \leq \kappa \leq 1/4$. Therefore, for all $\zeta \in [t_k, t_{k+1}]$, it holds that $e^{\zeta - t_k} + e^{-\zeta + t_k} < 2.1$, and as a consequence,

$$\begin{aligned}\|v_{c,t}\|_2 &\leq 12 \int_{t_k+\tau}^t (\sigma_{T-t_k}^{-2} + \lambda_{T-t_k}^2 \sigma_{T-t_k}^{-4}) d \|h_{t_k,\zeta}\|_2 \mathrm{d}\zeta \\ &\leq \int_{t_k+\tau}^t \frac{34(\sigma_{T-t_k}^{-2} + \lambda_{T-t_k}^2 \sigma_{T-t_k}^{-4}) d}{e^{\zeta - t_k} - e^{-\zeta + t_k}} \Big[\int_0^{\zeta - t_k} e^{\zeta - t_k - r} \big(\|B_{t_k+r} - B_{t_k}\|_2 + \gamma_{t_k, t_k+r} \big) \mathrm{d}r \Big] \mathrm{d}\zeta.\end{aligned} \tag{52}$$

We then proceed to upper bound the norms of $v_{a,t}$ and $v_{d,t}$. More specifically,

$$\|v_{a,t}\|_2 \leq \frac{1}{\sqrt{2}} (e^{t-t_k} + e^{-t+t_k} - 2) \big\| s(t_k, Y_{t_k}) \big\|_2 \leq \frac{(t-t_k)^2}{4} \big\| s(t_k, Y_{t_k}) \big\|_2, \tag{53}$$

$$\|v_{d,t}\|_2 \leq 2(t-t_k) \int_0^{t-t_k} \big\| s(t_k + r, Y_{t_k+r}) - s(t_k, Y_{t_k}) \big\|_2 \mathrm{d}r. \tag{54}$$

With regards to $v_{e,t}$, it follows from the triangle inequality and the assumption $\Delta_k \leq \kappa \leq 1/4$ that

$$\|v_{e,t}\|_2 \leq 10(t-t_k) d \big(\sigma_{T-t_k}^{-2} + \lambda_{T-t_k}^2 \sigma_{T-t_k}^{-4} \big) \int_0^{t-t_k} (\gamma_{t_k, t_k+r} + \|B_{t_k+r} - B_{t_k}\|_2) \mathrm{d}r. \tag{55}$$

To bound the norm of $v_{f,t}$, we find it helpful to bound the norm of the vector $s(\zeta, Y_\zeta) - s(t_k, Y_\zeta)$, towards which we resort to the third point of Lemma B.2. More precisely, for all $r \in [t_k, \zeta]$, the third point of Lemma B.2 tells us that

$$\partial_r s(r, Y_\zeta) = \frac{\lambda_{T-r} + \lambda_{T-r}^3}{\sigma_{T-r}^4} m(r, Y_\zeta) - \frac{2\lambda_{T-r}^2}{\sigma_{T-r}^4} Y_\zeta - \frac{\lambda_{T-r}}{\sigma_{T-r}^2} \mathbb{E}\big[\big(\theta - m(r, Y_\zeta) \big) \mathcal{F}(\theta, Y_\zeta, r) \mid \lambda_{T-r}\theta + \sigma_{T-r} g = Y_\zeta \big],$$

where $(\theta, g) \sim q_0 \otimes \mathcal{N}(0, I_d)$, and $m(r, y) = \mathbb{E}[\theta \mid \lambda_{T-r}\theta + \sigma_{T-r} g = y]$. Under Assumption A, it holds that $\|m(r, Y_\zeta)\|_2 \leq \sqrt{d}$ and

$$|\mathcal{F}(\theta, Y_\zeta, r)| \leq \lambda_{T-r}^2 \sigma_{T-r}^{-4} d + (\lambda_{T-r} + \lambda_{T-r}^3) \sigma_{T-r}^{-4} \sqrt{d} \|Y_\zeta\|_2.$$

Since $s(\zeta, Y_\zeta) - s(t_k, Y_\zeta) = \int_{t_k}^\zeta \partial_r s(r, Y_\zeta) \mathrm{d}r$, we see that for all $\zeta \in [t_k, t_{k+1}]$,

$$\begin{aligned}\big\| s(\zeta, Y_\zeta) - s(t_k, Y_\zeta) \big\|_2 &\leq \frac{\sqrt{d}(\zeta - t_k)(\lambda_{T-\zeta} + \lambda_{T-\zeta}^3)}{\sigma_{T-\zeta}^4} + \frac{2(\zeta - t_k)\lambda_{T-\zeta}^2 \|Y_\zeta\|_2}{\sigma_{T-\zeta}^4} \\ &+ \Big(\frac{d\lambda_{T-\zeta}^2}{\sigma_{T-\zeta}^4} + \frac{\sqrt{d}(\lambda_{T-\zeta} + \lambda_{T-\zeta}^3)\|Y_\zeta\|_2}{\sigma_{T-\zeta}^4} \Big) \cdot \int_{t_k}^\zeta \frac{\lambda_{T-r} \mathbb{E}[\|\theta - m(r, Y_\zeta)\|_2 \mid \lambda_{T-r}\theta + \sigma_{T-r} g = Y_\zeta]}{\sigma_{T-r}^2} \mathrm{d}r.\end{aligned} \tag{56}$$

Further, we make the observation that

$$\begin{aligned}&\|s(\zeta, Y_\zeta) - s(t_k, Y_{t_k}) - \sqrt{2} \nabla_x s(t_k, Y_{t_k} + h_{t_k,\zeta}) H_\zeta^\tau\|_2 \\ &\overset{(i)}{\leq} \|s(\zeta, Y_\zeta) - s(t_k, Y_\zeta)\|_2 + \|s(t_k, Y_\zeta) - s(t_k, Y_{t_k}) - \nabla_x s(t_k, Y_{t_k} + h_{t_k,\zeta})(Y_\zeta - Y_{t_k})\|_2 \\ &\quad + \|\nabla_x s(t_k, Y_{t_k} + h_{t_k,\zeta})(Y_\zeta - Y_{t_k} - \sqrt{2}B_\zeta + \sqrt{2}B_{t_k})\|_2 + \|\sqrt{2}\nabla_x s(t_k, Y_{t_k} + h_{t_k,\zeta})(B_\zeta - B_{t_k} - H_\zeta^\tau)\|_2 \\ &\overset{(ii)}{\leq} \|s(\zeta, Y_\zeta) - s(t_k, Y_\zeta)\|_2 + 8d(\sigma_{T-t_k}^{-2} + \lambda_{T-t_k}^2 \sigma_{T-t_k}^{-4}) \|Y_\zeta - Y_{t_k}\|_2 \\ &\quad + 4d(\sigma_{T-t_k}^{-2} + \lambda_{T-t_k}^2 \sigma_{T-t_k}^{-4}) \|Y_\zeta - Y_{t_k} - \sqrt{2}B_\zeta + \sqrt{2}B_{t_k}\|_2 + 6d(\sigma_{T-t_k}^{-2} + \lambda_{T-t_k}^2 \sigma_{T-t_k}^{-4}) \gamma_{t_k,\zeta}.\end{aligned} \tag{57}$$

In the above display, (i) comes from the triangle inequality, whereas (ii) is by the first point of Lemma B.2 and Assumption A. Substituting the upper bounds in Eqs. (56) and (57) into the definition of $v_{f,t}$ yields

$$
\begin{aligned}
&\|v_{f,t}\|_2 \\
&\le \sqrt{2}\int_{t_k+\tau}^{t}\Big(\frac{\sqrt{d}(\zeta-t_k)(\lambda_{T-\zeta}+\lambda_{T-\zeta}^3)}{\sigma_{T-\zeta}^4}+\frac{2(\zeta-t_k)\lambda_{T-\zeta}^2\|Y_\zeta\|_2}{\sigma_{T-\zeta}^4}\Big)\mathrm{d}\zeta \\
&\le \sqrt{2}\int_{t_k+\tau}^{t}\Big(\frac{d\lambda_{T-\zeta}^2}{\sigma_{T-\zeta}^4}+\frac{\sqrt{d}(\lambda_{T-\zeta}+\lambda_{T-\zeta}^3)\|Y_\zeta\|_2}{\sigma_{T-\zeta}^4}\Big)\cdot\int_{t_k}^{\zeta}\frac{\lambda_{T-r}\mathbb{E}[\|\theta-m(r,Y_\zeta)\|_2\mid\lambda_{T-r}\theta+\sigma_{T-r}g=Y_\zeta]}{\sigma_{T-r}^2}\mathrm{d}r\mathrm{d}\zeta \\
&+d\big(\sigma_{T-t_k}^{-2}+\lambda_{T-t_k}^2\sigma_{T-t_k}^{-4}\big)\int_{t_k+\tau}^{t}(9\gamma_{t_k,\zeta}+6\|Y_\zeta-Y_{t_k}-\sqrt{2}B_\zeta+\sqrt{2}B_{t_k}\|_2+12\|Y_\zeta-Y_{t_k}\|_2)\mathrm{d}\zeta. \qquad (58)
\end{aligned}
$$

The next step is to upper bound the norm of $v_{b,t}$, towards which we first make note of the following expression

$$
\begin{aligned}
v_{b,t}&=\int_{t_k+\tau}^{t}\int_{0}^{\zeta-t_k}\int_{0}^{e^{\zeta-t_k-r}}\mathrm{d}x(\mathrm{d}B_{t_k+r}-\mathrm{d}H_{t_k+r}^{\tau})\mathrm{d}\zeta \\
&=\int_{t_k+\tau}^{t}\int_{0}^{\zeta-t_k}\int_{0}^{e^{\zeta-t_k}}\mathbb{1}\{r\le\zeta-t_k-\log x\}\mathrm{d}x(\mathrm{d}B_{t_k+r}-\mathrm{d}H_{t_k+r}^{\tau})\mathrm{d}\zeta \\
&=\int_{t_k+\tau}^{t}\int_{0}^{e^{\zeta-t_k}}(B_{\zeta-\log x}-B_{t_k}-H_{\zeta-\log x}^{\tau})\mathrm{d}x\mathrm{d}\zeta.
\end{aligned}
$$

As a consequence, we have

$$
\|v_{b,t}\|_2\le\int_{t_k+\tau}^{t}\int_{0}^{e^{\zeta-t_k}}\gamma_{t_k,\zeta-\log x}\mathrm{d}x\mathrm{d}\zeta. \qquad (59)
$$

Finally, we can conclude the proof of the lemma using the upper bounds derived above. For $t\in[t_k,t_{k+1}]$, we define

$$
\gamma_*(t)=\sup_{s\in[t_k,t]}\gamma_{t_k,s}.
$$

We see that at $t=t_k$ we have $\gamma_*(t_k)=0$, and our goal is to upper bound $\gamma_*(t_{k+1})$. In addition, since the processes $(B_t-B_{t_k})_{t_k\le t\le t_{k+1}}$ and $(H_t^\tau)_{t_k\le t\le t_{k+1}}$ have continuous sample paths, the mapping $t\mapsto\gamma_*(t)$ is continuous on $t\in[t_k,t_{k+1}]$. Substitution of Eqs. (52) to (55), (58) and (59) into Eq. (51) gives

$$
\gamma_*(t)\le C_0+C_1\gamma_*(t), \qquad (60)
$$

where $C_0,C_1\in\mathbb{R}_{>0}$ are defined as follows:

$$
\begin{aligned}
C_0&=\frac{(t-t_k)^2}{4}\|s(t_k,Y_{t_k})\|_2+\int_{t_k+\tau}^{t}\frac{44(\sigma_{T-t_k}^{-2}+\lambda_{T-t_k}^2\sigma_{T-t_k}^{-4})d}{e^{\zeta-t_k}-e^{-\zeta+t_k}}\int_{0}^{\zeta-t_k}\|B_{t_k+r}-B_{t_k}\|_2\mathrm{d}r\mathrm{d}\zeta \\
&+2(t-t_k)\int_{0}^{t-t_k}\big\|s(t_k+r,Y_{t_k+r})-s(t_k,Y_{t_k})\big\|_2\mathrm{d}r+10(t-t_k)d(\sigma_{T-t_k}^{-2}+\lambda_{T-t_k}^2\sigma_{T-t_k}^{-4})\int_{0}^{t-t_k}\|B_{t_k+r}-B_{t_k}\|_2\mathrm{d}r \\
&+2\sqrt{2}\int_{t_k+\tau}^{t}\Big(\frac{d\lambda_{T-\zeta}^2}{\sigma_{T-\zeta}^4}+\frac{\sqrt{d}(\lambda_{T-\zeta}+\lambda_{T-\zeta}^3)\|Y_\zeta\|_2}{\sigma_{T-\zeta}^4}\Big)\cdot\int_{t_k}^{\zeta}\frac{\lambda_{T-r}\sqrt{d}}{\sigma_{T-r}^2}\mathrm{d}r\mathrm{d}\zeta \\
&+\sqrt{2}\int_{t_k+\tau}^{t}\Big(\frac{\sqrt{d}(\zeta-t_k)(\lambda_{T-\zeta}+\lambda_{T-\zeta}^3)}{\sigma_{T-\zeta}^4}+\frac{2(\zeta-t_k)\lambda_{T-\zeta}^2\|Y_\zeta\|_2}{\sigma_{T-\zeta}^4}\Big)\mathrm{d}\zeta \\
&+d(\sigma_{T-t_k}^{-2}+\lambda_{T-t_k}^2\sigma_{T-t_k}^{-4})\int_{t_k+\tau}^{t}(6\|Y_\zeta-Y_{t_k}-\sqrt{2}B_\zeta+\sqrt{2}B_{t_k}\|_2+12\|Y_\zeta-Y_{t_k}\|_2)\mathrm{d}\zeta, \\
C_1&=1.3(t-t_k)+(53(t-t_k)+10(t-t_k)^2)(\sigma_{T-t_k}^{-2}+\lambda_{T-t_k}^2\sigma_{T-t_k}^{-4})d.
\end{aligned}
$$

Under Assumption B, we know that $C_1 \le 1/2$ for all $t \in [t_k, t_{k+1}]$, and hence $\gamma_*(t) \le 2C_0$. Invoking Lemma A.3 tells us that: for all $\zeta \in [t_k, t_{k+1}]$ we have

$$\begin{aligned}
&\mathbb{E}\big[\|Y_\zeta - Y_{t_k}\|_2^2\big] \lesssim d(\zeta - t_k) + d\sigma_{T-\zeta}^{-2}(\zeta - t_k)^2, \qquad \mathbb{E}\big[\|Y_\zeta\|_2^2\big] \lesssim d, \qquad \mathbb{E}\big[\|B_\zeta - B_{t_k}\|_2^2\big] \lesssim d(\zeta - t_k),\\
&\mathbb{E}\big[\|s(\zeta, Y_\zeta)\|_2^2\big] \lesssim d\sigma_{T-\zeta}^{-2}, \qquad \mathbb{E}\big[\|Y_\zeta - Y_{t_k} - \sqrt{2}B_\zeta + \sqrt{2}B_{t_k}\|_2^2\big] \lesssim d\sigma_{T-\zeta}^{-2}(\zeta - t_k)^2,\\
&\mathbb{E}\big[\|Y_\zeta - Y_{t_k}\|_2^4\big] \lesssim d^2(\zeta - t_k)^2 + d^2\sigma_{T-\zeta}^{-4}(\zeta - t_k)^4, \qquad \mathbb{E}\big[\|Y_\zeta\|_2^4\big] \lesssim d^2, \qquad \mathbb{E}\big[\|B_\zeta - B_{t_k}\|_2^4\big] \lesssim d^2(\zeta - t_k)^2,\\
&\mathbb{E}\big[\|s(\zeta, Y_\zeta)\|_2^2\big] \lesssim d^2\sigma_{T-\zeta}^{-4}, \qquad \mathbb{E}\big[\|Y_\zeta - Y_{t_k} - \sqrt{2}B_\zeta + \sqrt{2}B_{t_k}\|_2^2\big] \lesssim d^2\sigma_{T-\zeta}^{-4}(\zeta - t_k)^4.
\end{aligned}$$

Taking the expectation of C_0^2 and C_0^4 implies that for all $t \in [t_k, t_{k+1}]$,

$$\begin{aligned}
\mathbb{E}[\gamma_*(t)^2] &\lesssim \sigma_{T-t}^{-2}(t - t_k)^4 d + (\sigma_{T-t_k}^{-2} + \lambda_{T-t_k}^2\sigma_{T-t_k}^{-4})^2(t - t_k)^3 d^3 + \lambda_{T-t}^4\sigma_{T-t}^{-12}(t - t_k)^4 d^3 + \lambda_{T-t}^2\sigma_{T-t}^{-8}(t - t_k)^4 d,\\
\mathbb{E}[\gamma_*(t)^4] &\lesssim \sigma_{T-t}^{-4}(t - t_k)^8 d^2 + (\sigma_{T-t_k}^{-2} + \lambda_{T-t_k}^2\sigma_{T-t_k}^{-4})^4(t - t_k)^6 d^6 + \lambda_{T-t}^8\sigma_{T-t}^{-24}(t - t_k)^8 d^6 + \lambda_{T-t}^4\sigma_{T-t}^{-16}(t - t_k)^8 d^2.
\end{aligned}$$

The proof is thus complete.

C.3 Proof of Lemma 4.3

According to Lemma 4.1, we see that: in order to upper bound $\mathsf{KL}(Q_{T-\delta} \,\|\, \overline{Q}_{T-\delta})$, it suffices to control

$$\sum_{k=0}^{K-1} \int_{t_k}^{t_{k+1}} \mathbb{E}\Big[\big\|\mathcal{F}(t, Y_t) - \overline{\mathcal{F}}(t, Y_{t_k}, (H_s^\tau)_{t_k \le s \le t})\big\|_2^2\Big] \mathrm{d}t.$$

In view of Lemma B.2 and Assumption A, we know that for all $t \in [0, T)$ and $x \in \mathbb{R}^d$, it holds that

$$\|\nabla_x s(t, x)\|_2 \lesssim (\sigma_{T-t}^{-2} + \lambda_{T-t}^2\sigma_{T-t}^{-4})d, \qquad \|\nabla_x^2 s(t, x)\|_2 \lesssim \lambda_{T-t}^3\sigma_{T-t}^{-6}d^{3/2},$$

which in turn allow one to show that $\overline{\mathcal{F}}$ is Lipschitz continuous with respect to the last input. More precisely,

$$\begin{aligned}
&\big\|\overline{\mathcal{F}}(t, Y_{t_k}, (H_s^\tau)_{t_k \le s \le t}) - \overline{\mathcal{F}}(t, Y_{t_k}, (B_s - B_{t_k})_{t_k \le s \le t})\big\|_2\\
&\lesssim \|\nabla_x s\big(t_k, \overline{Y}_{t_k} + h_{t_k,t}\big)\|_2 \|B_t - B_{t_k} - H_t^\tau\|_2 + \|\nabla_x s\big(t_k, \overline{Y}_{t_k} + h_{t_k,t}\big) - \nabla_x s\big(t_k, \overline{Y}_{t_k} + b_{t_k,t}\big)\|_2 \cdot \|B_t - B_{t_k}\|_2\\
&+ \|\nabla_x s\big(t_k, \overline{Y}_{t_k} + h_{t_k,t}\big)\|_2 \|h_{t_k,t} - b_{t_k,t}\|_2 + \|\nabla_x s\big(t_k, \overline{Y}_{t_k} + h_{t_k,t}\big) - \nabla_x s\big(t_k, \overline{Y}_{t_k} + b_{t_k,t}\big)\|_2 \cdot \|b_{t_k,t}\|_2\\
&\lesssim \Big(d(\sigma_{T-t}^{-2} + \lambda_{T-t}^2\sigma_{T-t}^{-4}) + d^{3/2}\lambda_{T-t}^3\sigma_{T-t}^{-6}(\|B_t - B_{t_k}\|_2 + \|b_{t_k,t}\|_2)\Big) \sup_{t_k \le t \le t_{k+1}} \big\|B_t - B_{t_k} - H_t^\tau\big\|_2,
\end{aligned}$$

with

$$\begin{aligned}
b_{t_k,t} &= 2\sqrt{2}(e^{t-t_k} - e^{-t+t_k})^{-1} \int_0^{t-t_k} e^{t-t_k-r}(B_{t_k+r} - B_{t_k})\mathrm{d}r,\\
h_{t_k,t} &= 2\sqrt{2}(e^{t-t_k} - e^{-t+t_k})^{-1} \int_0^{t-t_k} e^{t-t_k-r} H_{t_k+r}^\tau \mathrm{d}r.
\end{aligned}$$

Putting the above inequality together with Lemma 4.2 gives

$$\begin{aligned}
&\sum_{k=0}^{K-1} \int_{t_k}^{t_{k+1}} \mathbb{E}\Big[\big\|\mathcal{F}(t, Y_t) - \overline{\mathcal{F}}(t, Y_{t_k}, (H_s^\tau)_{t_k \le s \le t})\big\|_2^2\Big] \mathrm{d}t\\
&\lesssim \sum_{k=0}^{K-1} \int_{t_k}^{t_{k+1}} \mathbb{E}\Big[\big\|\mathcal{F}(t, Y_t) - \overline{\mathcal{F}}(t, Y_{t_k}, (B_s - B_{t_k})_{t_k \le s \le t})\big\|_2^2\Big] \mathrm{d}t\\
&\quad + \sum_{k=0}^{K-1} \int_{t_k}^{t_{k+1}} \mathbb{E}\Big[\big\|\overline{\mathcal{F}}(t, Y_{t_k}, (H_s^\tau)_{t_k \le s \le t}) - \overline{\mathcal{F}}(t, Y_{t_k}, (B_s - B_{t_k})_{t_k \le s \le t})\big\|_2^2\Big] \mathrm{d}t\\
&\lesssim \sum_{k=0}^{K-1} \int_{t_k}^{t_{k+1}} \mathbb{E}\Big[\big\|\mathcal{F}(t, Y_t) - \overline{\mathcal{F}}(t, Y_{t_k}, (B_s - B_{t_k})_{t_k \le s \le t})\big\|_2^2\Big] \mathrm{d}t
\end{aligned}$$

$$+\sum_{k=0}^{K-1}\Delta_k\Big(d^2(\sigma_{T-t}^{-2}+\lambda_{T-t}^2\sigma_{T-t}^{-4})^2+d^3\lambda_{T-t}^6\sigma_{T-t}^{-12}\Delta_k\Big)$$
$$\times\Big(\sigma_{T-t_{k+1}}^{-2}\Delta_k^4 d+(\sigma_{T-t_k}^{-2}+\lambda_{T-t_k}^2\sigma_{T-t_k}^{-4})^2\Delta_k^3 d^3+\lambda_{T-t_{k+1}}^4\sigma_{T-t_{k+1}}^{-12}\Delta_k^4 d^3+\lambda_{T-t_{k+1}}^2\sigma_{T-t_{k+1}}^{-8}\Delta_k^4 d\Big).$$

It is seen from the triangle inequality that

$$\sum_{k=0}^{K-1}\int_{t_k}^{t_{k+1}}\mathbb{E}\Big[\big\|\mathcal{F}(t,Y_t)-\overline{\mathcal{F}}(t,Y_{t_k},(B_s-B_{t_k})_{t_k\le s\le t})\big\|_2^2\Big]\mathrm{d}t\lesssim\mathbb{T}_1+\mathbb{T}_2+\mathbb{T}_3+\mathbb{T}_4,$$

where

$$\mathbb{T}_1=\sum_{k=0}^{K-1}\int_{t_k}^{t_{k+1}}(e^{t-t_k}-e^{-t+t_k})^2\mathbb{E}\big[\|s(t_k,Y_{t_k})\|_2^2\big]\mathrm{d}t,$$
$$\mathbb{T}_2=\sum_{k=0}^{K-1}\int_{t_k}^{t_{k+1}}\mathbb{E}\big[\big\|s(t,Y_t)-s(t_k,Y_{t_k})-\sqrt{2}\nabla_x s(t_k,Y_{t_k}+b_{t_k,t})(B_t-B_{t_k})\big\|_2^2\big]\mathrm{d}t,$$
$$\mathbb{T}_3=\sum_{k=0}^{K-1}\int_{t_k}^{t_{k+1}}\int_0^{t-t_k}\mathbb{E}\big[\|s(t_k+r,Y_{t_k+r})-s(t_k,Y_{t_k})-\sqrt{2}\nabla_x s(t_k,Y_{t_k}+b_{t_k,t})(B_{t_k+r}-B_{t_k})\|_2^2\big]\mathrm{d}r\mathrm{d}t,$$
$$\mathbb{T}_4=\sum_{k=0}^{K-1}\int_{t_k}^{t_{k+1}}\mathbb{E}\big[\|s(t_k,Y_{t_k}+b_{t_k,t})-s(t_k,Y_{t_k})-\nabla_x s(t_k,Y_{t_k}+b_{t_k,t})b_{t_k,t}\|_2^2\big]\mathrm{d}t.$$

The terms $\mathbb{T}_1$, $\mathbb{T}_2$, $\mathbb{T}_3$ and $\mathbb{T}_4$ are upper bounded separately in Lemma C.1 below, whose proof can be found in Appendix C.5.

Lemma C.1. *Under the assumptions of Lemma 4.3, it holds that*

1. $\mathbb{T}_1\lesssim\sum_{k=0}^{K-1}\sigma_{T-t_k}^{-2}\Delta_k^3 d$;
2. $\mathbb{T}_2\lesssim\sum_{k=0}^{K-1}\big(d^3\sigma_{T-t_{k+1}}^{-6}\Delta_k^3+d^7\lambda_{T-t_k}^8\sigma_{T-t_k}^{-16}\Delta_k^4\big)$;
3. $\mathbb{T}_3\lesssim\sum_{k=0}^{K-1}\big(d^3\sigma_{T-t_{k+1}}^{-6}\Delta_k^4+d^7\lambda_{T-t_k}^8\sigma_{T-t_k}^{-16}\Delta_k^5\big)$;
4. $\mathbb{T}_4\lesssim\sum_{k=0}^{K-1}\big(d^3\sigma_{T-t_{k+1}}^{-6}\Delta_k^3+d^7\lambda_{T-t_k}^8\sigma_{T-t_k}^{-16}\Delta_k^4\big)$.

As a consequence of Lemma C.1, we reach

$$\mathsf{KL}(Q_{T-\delta}\,\|\,\overline{Q}_{T-\delta})\lesssim\sum_{k=0}^{K-1}\big(d^3\sigma_{T-t_{k+1}}^{-6}\Delta_k^3+d^7\lambda_{T-t_k}^8\sigma_{T-t_k}^{-16}\Delta_k^4\big).$$

Let $K_0=\inf\{k:T-t_k\le 1\}$, look at the indices that are above and below K_0 separately. Specifically, for all $k\le K_0-1$ we have $\sigma_{T-t_k}^{-2}\lesssim 1$. In addition, for all $K-1\ge k\ge K_0$ we have $\sigma_{T-t_k}^{-2}\lesssim(T-t_k)^{-1}$, and hence under Assumption B we have $d^3\sigma_{T-t_{k+1}}^{-6}\Delta_k^3+d^7\lambda_{T-t_k}^8\sigma_{T-t_k}^{-16}\Delta_k^4\lesssim d^3\kappa^2\Delta_k+d^7\lambda_{T-t_k}^8\sigma_{T-t_k}^{-4}\kappa^3\Delta_k$. Putting these together yields

$$\begin{aligned}&\sum_{k=0}^{K-1}\big(d^3\sigma_{T-t_{k+1}}^{-6}\Delta_k^3+d^7\lambda_{T-t_k}^8\sigma_{T-t_k}^{-16}\Delta_k^4\big)\\&=\sum_{k=0}^{K_0-1}\big(d^3\sigma_{T-t_{k+1}}^{-6}\Delta_k^3+d^7\lambda_{T-t_k}^8\sigma_{T-t_k}^{-16}\Delta_k^4\big)+\sum_{k=K_0}^{K-1}\big(d^3\kappa^2\Delta_k+d^7\lambda_{T-t_k}^8\sigma_{T-t_k}^{-4}\kappa^3\Delta_k\big)\\&\lesssim d^3\kappa^2T+d^7\kappa^3(\delta^{-1}+T),\end{aligned}$$

as claimed in the lemma.

C.4 Proof of Lemma 4.4

Denoting by $\overline{Q}^{\mathsf{dis}}_{T-\delta}$ the distribution of $(\overline{Y}_{t_k})_{0\le k\le K}$, we observe that

$$\mathsf{KL}(Q^{\mathsf{dis}}_{T-\delta} \,\|\, \widehat{Q}^{\mathsf{dis}}_{T-\delta}) = \underbrace{\int \mathrm{d}Q^{\mathsf{dis}}_{T-\delta} \log \frac{\mathrm{d}Q^{\mathsf{dis}}_{T-\delta}}{\mathrm{d}\overline{Q}^{\mathsf{dis}}_{T-\delta}}}_{\text{(i)}} + \underbrace{\int \mathrm{d}Q^{\mathsf{dis}}_{T-\delta} \log \frac{\mathrm{d}\overline{Q}^{\mathsf{dis}}_{T-\delta}}{\mathrm{d}\widehat{Q}^{\mathsf{dis}}_{T-\delta}}}_{\text{(ii)}}, \tag{61}$$

leaving us with two terms to control.

Bounding the term (i)

Note that the term (i) is essentially (i) $= \mathsf{KL}(Q^{\mathsf{dis}}_{T-\delta} \,\|\, \overline{Q}^{\mathsf{dis}}_{T-\delta})$. Recall that the data processing inequality (Polyanskiy, 2020, Theorem 7.2) asserts that $\mathsf{KL}\big(P_{f(X)} \,\|\, P_{f(Y)}\big) \le \mathsf{KL}(P_X \,\|\, P_Y)$ holds for any two random objects X and Y on the same space and any mapping f, with P_Z the distribution of Z. Taking this together with Lemma 4.3, we reach

$$\int \mathrm{d}Q^{\mathsf{dis}}_{T-\delta} \log \frac{\mathrm{d}Q^{\mathsf{dis}}_{T-\delta}}{\mathrm{d}\overline{Q}^{\mathsf{dis}}_{T-\delta}} \le \int \mathrm{d}Q_{T-\delta} \log \frac{\mathrm{d}Q_{T-\delta}}{\mathrm{d}\overline{Q}_{T-\delta}} \lesssim d^3\kappa^2 T + d^7\kappa^3(\delta^{-1}+T).$$

Bounding the term (ii)

Recall that $\widehat{Q}_{T-\delta}$ is the distribution of process (10a) and $\overline{Q}_{T-\delta}$ is that of process (21). By Eq. (10a) and Eq. (21), we see that for each $k \in \{0,1,\dots,K-1\}$,

$$\begin{aligned}
\overline{Y}_{t_{k+1}} &= e^{\Delta_k}\overline{Y}_{t_k} + (e^{\Delta_k} - e^{-\Delta_k}) s(t_k, \overline{Y}_{t_k} + g_{t_k,t_{k+1}}) + \sqrt{2}\int_0^{\Delta_k} e^{\Delta_k - r}\mathrm{d}W_{t_k+r}, \\
\widehat{Y}_{t_{k+1}} &= e^{\Delta_k}\widehat{Y}_{t_k} + (e^{\Delta_k} - e^{-\Delta_k}) \widehat{s}(t_k, \widehat{Y}_{t_k} + g_{t_k,t_{k+1}}) + \sqrt{2}\int_0^{\Delta_k} e^{\Delta_k - r}\mathrm{d}W_{t_k+r},
\end{aligned} \tag{62}$$

where we recall that $g_{t_k,t_{k+1}}$ is defined in Eq. (10b). Note that $g_{t_k,t_{k+1}}$ and $\sqrt{2}\int_0^{\Delta_k} e^{\Delta_k - r}\mathrm{d}W_{t_k+r}$ are correlated Gaussian random vectors, which admit simpler expressions. Specifically, from Lemma A.1, we can write

$$g_{t_k,t_{k+1}} = \zeta_{k,1} g_{k,1}, \qquad \sqrt{2}\int_0^{\Delta_k} e^{\Delta_k - r}\mathrm{d}W_{t_k+r} = \zeta_{k,2} g_{k,1} + \zeta_{k,3} g_{k,2}, \tag{63}$$

where $g_{k,1}, g_{k,2} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$, and

$$\zeta_{k,1} = \frac{2\sqrt{2} f_1(\Delta_k)^{1/2}}{e^{\Delta_k} - e^{-\Delta_k}}, \qquad \zeta_{k,2} = \frac{\sqrt{2} f_3(\Delta_k)}{f_1(\Delta_k)^{1/2}}, \qquad \zeta_{k,3} = \sqrt{2 f_2(\Delta_k) - \frac{2 f_3(\Delta_k)^2}{f_1(\Delta_k)}}. \tag{64}$$

In the above display, we recall the definitions of f_1, f_2 and f_3 in Eq. (32).

For every $0 \le k \le K$, denote by Q_k the distribution of $(Y_{t_0}, Y_{t_1}, \cdots, Y_{t_k})$, $\overline{Q}_k$ the distribution of $(\overline{Y}_{t_0}, \overline{Y}_{t_1}, \cdots, \overline{Y}_{t_k})$, and $\widehat{Q}_k$ that of $(\widehat{Y}_{t_0}, \widehat{Y}_{t_1}, \cdots, \widehat{Y}_{t_k})$. Therefore, it follows from Eq. (62) that

$$\begin{aligned}
&\int \mathrm{d}Q^{\mathsf{dis}}_{T-\delta} \log \frac{\mathrm{d}\overline{Q}^{\mathsf{dis}}_{T-\delta}}{\mathrm{d}\widehat{Q}^{\mathsf{dis}}_{T-\delta}} \\
&= \sum_{k=0}^{K-1} \int \mathrm{d}Q_{k+1} \log \frac{\mathrm{d}\overline{Q}_{k+1}(Y_{t_{k+1}} \mid Y_{t_0}, \cdots, Y_{t_k})}{\mathrm{d}\widehat{Q}_{k+1}(Y_{t_{k+1}} \mid Y_{t_0}, \cdots, Y_{t_k})} \\
&= \sum_{k=0}^{K-1} \int \mathrm{d}Q_{k+1} \log \frac{\int \exp\big(-\|Y_{t_{k+1}} - e^{\Delta_k} Y_{t_k} - (e^{\Delta_k} - e^{-\Delta_k}) s(t_k, Y_{t_k} + \zeta_{k,1} g) - \zeta_{k,2} g\|_2^2 / (2\zeta_{k,3}^2)\big)\phi(g)\mathrm{d}g}{\int \exp\big(-\|Y_{t_{k+1}} - e^{\Delta_k} Y_{t_k} - (e^{\Delta_k} - e^{-\Delta_k}) \widehat{s}(t_k, Y_{t_k} + \zeta_{k,1} g) - \zeta_{k,2} g\|_2^2 / (2\zeta_{k,3}^2)\big)\phi(g)\mathrm{d}g},
\end{aligned}$$

where $\phi(\cdot)$ denotes the probability density function of $\mathcal{N}(0, I_d)$. Recalling that $s(t,x) = \sigma_{T-t}^{-2}(-x+\lambda_{T-t}m(t,x))$ and $\widehat{s}(t,x) = \sigma_{T-t}^{-2}(-x+\lambda_{T-t}\widehat{m}(t,x))$ (with m and $\widehat{m}$ introduced in Remark 3.3), we can further deduce that

$$\begin{aligned}
&\int \mathrm{d}Q_{T-\delta}^{\mathsf{dis}} \log \frac{\mathrm{d}\overline{Q}_{T-\delta}^{\mathsf{dis}}}{\mathrm{d}\widehat{Q}_{T-\delta}^{\mathsf{dis}}} \qquad (65)\\
&= \sum_{k=0}^{K-1} \int \mathrm{d}Q_{k+1} \log \frac{\int \exp\big(\eta_k\langle m(t_k, Y_{t_k}+\zeta_{k,1}g), v_k-\kappa_k g\rangle - \gamma_k\|m(t_k,Y_{t_k}+\zeta_{k,1}g)\|_2^2/2\big)\phi_{\xi_k,\tau_k}(g)\mathrm{d}g}{\int \exp\big(\eta_k\langle \widehat{m}(t_k, Y_{t_k}+\zeta_{k,1}g), v_k-\kappa_k g\rangle - \gamma_k\|\widehat{m}(t_k,Y_{t_k}+\zeta_{k,1}g)\|_2^2/2\big)\phi_{\xi_k,\tau_k}(g)\mathrm{d}g},
\end{aligned}$$

where $\phi_{\xi,\tau}(\cdot)$ denotes the probability density function for $\mathcal{N}(\xi, \tau^2 I_d)$, and we set

$$\begin{aligned}
\eta_k &= \zeta_{k,3}^{-2}\sigma_{T-t_k}^{-2}\lambda_{T-t_k}(e^{\Delta_k}-e^{-\Delta_k}),\\
v_k &= Y_{t_{k+1}} + \big(\sigma_{T-t_k}^{-2}(e^{\Delta_k}-e^{-\Delta_k}) - e^{\Delta_k}\big)Y_{t_k},\\
\kappa_k &= \zeta_{k,2} - \sigma_{T-t_k}^{-2}(e^{\Delta_k}-e^{-\Delta_k})\zeta_{k,1},\\
\gamma_k &= \zeta_{k,3}^{-2}\sigma_{T-t_k}^{-4}\lambda_{T-t_k}^2(e^{\Delta_k}-e^{-\Delta_k})^2,\\
\tau_k^2 &= (\zeta_{k,3}^2+\kappa_k^2)^{-1}\zeta_{k,3}^2,\\
\xi_k &= \zeta_{k,3}^{-2}\tau_k^2\kappa_k v_k.
\end{aligned} \qquad (66)$$

We find it helpful to single out the following useful upper bounds (the proofs are omitted as they follow from straightforward calculus techniques):

$$\eta_k \lesssim \sigma_{T-t_k}^{-2}\lambda_{T-t_k}, \qquad |\kappa_k| \lesssim \Delta_k^{1/2}, \qquad \gamma_k \lesssim \kappa\lambda_{T-t_k}^2, \qquad \tau_k^2 \lesssim 1. \qquad (67)$$

In addition, it can be verified that: if $g \sim \mathcal{N}(\xi_k, \tau_k^2 I_d)$, then one can write

$$Y_{t_k}+\zeta_{k,1}g = Y_{t_k} + \zeta_{k,1}\zeta_{k,3}^{-2}\tau_k^2\kappa_k\big(Y_{t_{k+1}} - e^{\Delta_k}Y_{t_k} + \sigma_{T-t_k}^{-2}(e^{\Delta_k}-e^{-\Delta_k})Y_{t_k}\big) + \zeta_{k,1}\tau_k g', \qquad (68)$$

where $g' \sim \mathcal{N}(0, I_d)$ is independent of $(Y_{t_k}, Y_{t_{k+1}})$.

Now let us analyze the quantities $\zeta_{k,1}, \zeta_{k,2}, \zeta_{k,3}$ and those defined in Eq. (66) in the lemma below.

Lemma C.2. *Under the assumptions of Lemma 4.4, it holds that*

$$\begin{aligned}
|\zeta_{k,1}\zeta_{k,3}^{-2}\tau_k^2\kappa_k| &\le 0.65,\\
\sigma_{T-t_k}^{-2}\big(1-e^{-2\Delta_k}\big) &\le 3.2\sqrt{\Delta_k\kappa},\\
\sigma_{T-t_k}^{-2}\big(e^{\Delta_k}-e^{-\Delta_k}\big) &\le 3.25\sqrt{\Delta_k\kappa}.
\end{aligned}$$

Proof of Lemma C.2. Recall that by Assumption B we have $\Delta_k \le \kappa < 1/4$. As a consequence, we have $1-e^{-2\Delta_k} \le 2\Delta_k$ and $e^{\Delta_k}-e^{-\Delta_k} \le 81\Delta_k/40$.

If $T-t_k \ge 1/2$, then $\sigma_{T-t_k}^{-2} \le (1-e^{-1})^{-1}$. In this case, we have $\sigma_{T-t_k}^{-2}(1-e^{-2\Delta_k}) \le 3.2\Delta_k$, and $\sigma_{T-t_k}^{-2}(e^{\Delta_k}-e^{-\Delta_k}) \le 3.25\Delta_k$. On the other hand, if $T-t_k < 1/2$, then $\sigma_{T-t_k}^{-2} \le 0.8(T-t_k)^{-1}$, hence $\sigma_{T-t_k}^{-2}(1-e^{-2\Delta_k}) \le 1.6\Delta_k/(T-t_k) \le 1.6\sqrt{\kappa\Delta_k}$ and $\sigma_{T-t_k}^{-2}(e^{\Delta_k}-e^{-\Delta_k}) \le 1.62\Delta_k/(T-t_k) \le 1.62\sqrt{\kappa\Delta_k}$. This establishes the second and the third inequalities.

As for the first inequality, observe that $|\zeta_{k,1}\zeta_{k,3}^{-2}\tau_k^2\kappa_k| = |\zeta_{k,1}\kappa_k/(\zeta_{k,3}^2+\kappa_k^2)| \le |\zeta_{k,1}\zeta_{k,3}^{-1}|/2$. When $\Delta_k \le 1/4$, it holds that $\zeta_{k,1} \in [0.8\Delta_k^{1/2}, 0.9\Delta_k^{1/2}]$ and $\zeta_{k,3} \in [0.7\Delta_k^{1/2}, 0.75\Delta_k^{1/2}]$. Therefore, we have $|\zeta_{k,1}\zeta_{k,3}^{-2}\tau_k^2\kappa_k| \le |\zeta_{k,1}\zeta_{k,3}^{-1}|/2 \le 0.65$. □

Denoting by ν_k the marginal distribution of the random vector in Eq. (68), we provide an important property about ν_k in the next lemma.

Lemma C.3. *Under the assumptions of Lemma 4.4, it holds that* $\nu_k \overset{\mathrm{d}}{=} a_k Y_{t_{k+1}} + b_k g$, *where* a_k, b_k *are quantities satisfying* $|a_k - 1| \le 3.1\sqrt{\Delta_k \kappa}$ *and* $|b_k| \le 3.5\sqrt{\Delta_k}$. *Here,* $g \sim \mathcal{N}(0, I_d)$ *is independent of* $Y_{t_{k+1}}$, *and we recall that* $Y_{t_{k+1}} \sim q_{T-t_{k+1}}$.

Proof of Lemma C.3. Note that we can write $Y_{t_k} = e^{-\Delta_k} Y_{t_{k+1}} + \sqrt{1 - e^{-2\Delta_k}} G$, where $G \sim \mathcal{N}(0, I_d)$ is independent of $Y_{t_{k+1}}$. Substituting this equation into Eq. (68) yields

$$\begin{aligned}
\nu_k &\overset{\mathrm{d}}{=} \big(e^{-\Delta_k} + \sigma_{T-t_k}^{-2}(1 - e^{-2\Delta_k})\zeta_{k,1}\zeta_{k,3}^{-2}\tau_k^2\kappa_k\big) Y_{t_{k+1}} + \zeta_{k,1}\zeta_{k,3}^{-2}\tau_k^2\kappa_k\sqrt{1 - e^{-2\Delta_k}}\big(\sigma_{T-t_k}^{-2}(e^{\Delta_k} - e^{-\Delta_k}) - e^{\Delta_k}\big) G \\
&\quad + \sqrt{1 - e^{-2\Delta_k}} G + \zeta_{k,1}\tau_k g' \\
&\overset{\mathrm{d}}{=} a_k Y_{t_{k+1}} + b_k G,
\end{aligned}$$

where

$$\begin{aligned}
a_k &= e^{-\Delta_k} + \sigma_{T-t_k}^{-2}(1 - e^{-2\Delta_k})\zeta_{k,1}\zeta_{k,3}^{-2}\tau_k^2\kappa_k, \\
b_k &= \sqrt{\zeta_{k,1}^2\tau_k^2 + (1 - e^{-2\Delta_k})\big(\zeta_{k,1}\zeta_{k,3}^{-2}\tau_k^2\kappa_k(\sigma_{T-t_k}^{-2}(e^{\Delta_k} - e^{-\Delta_k}) - e^{\Delta_k}) + 1\big)^2}.
\end{aligned}$$

Note that $|1 - e^{-2\Delta_k}| \le 2\Delta_k \le 2\kappa$ and $\tau_k^2 \le 1$. Using these upper bounds and Lemma C.2, we reach

$$\begin{aligned}
|a_k - 1| &\le |e^{-\Delta_k} - 1| + \sigma_{T-t_k}^{-2}(1 - e^{-2\Delta_k})|\zeta_{k,1}\zeta_{k,3}^{-2}\tau_k^2\kappa_k| \le 3.1\sqrt{\Delta_k\kappa} \\
|b_k| &\le \sqrt{0.9^2\Delta_k + 2\Delta_k\big(1 + 0.65 \times (e^{1/4} + 3.25\kappa)\big)^2} \le 3.5\Delta_k^{1/2}
\end{aligned}$$

as claimed. □

By virtue of Assumption C and Lemma C.3, we know that

$$\mathbb{E}_{y\sim\nu_k}\big[\|m(t_k, y) - \widehat{m}(t_k, y)\|_2^2\big] = \sigma_{T-t_k}^4\lambda_{T-t_k}^{-2}\mathbb{E}_{y\sim\nu_k}\big[\|s(t_k, y) - \widehat{s}(t_k, y)\|_2^2\big] \le \sigma_{T-t_k}^4\lambda_{T-t_k}^{-2}\varepsilon_{\mathsf{score},k}^2.$$

In the sequel, we make the convention that conditional on $(Y_{t_k}, Y_{t_{k+1}})$, $g \sim \mathcal{N}(\xi_k, \tau_k^2 I_d)$, where we recall that (ξ_k, τ_k^2) are defined in Eq. (66). Note that ξ_k is a function of $(Y_{t_k}, Y_{t_{k+1}})$. For $y_k, y_{k+1} \in \mathbb{R}^d$, we define

$$p_k(y_k, y_{k+1}) = \mathbb{P}\big(\|m(t_k, Y_{t_k} + \zeta_{k,1}g) - \widehat{m}(t_k, Y_{t_k} + \zeta_{k,1}g)\|_2 \ge \sigma_{T-t_k}\lambda_{T-t_k}^{-1/2}\varepsilon_{\mathsf{score},k}^{1/2} \mid Y_{t_k} = y_k, Y_{t_{k+1}} = y_{k+1}\big).$$

Then by Chebyshev's inequality, one has

$$\mathbb{E}[p_k(Y_{t_k}, Y_{t_{k+1}})] \le \sigma_{T-t_k}^2\lambda_{T-t_k}^{-1}\varepsilon_{\mathsf{score},k}.$$

Conditioning on $(Y_{t_k}, Y_{t_{k+1}})$, we introduce the conditional event

$$\mathcal{S}_{Y_{t_k}, Y_{t_{k+1}}} = \Big\{g : \|m(t_k, Y_{t_k} + \zeta_{k,1}g) - \widehat{m}(t_k, Y_{t_k} + \zeta_{k,1}g)\|_2 \ge \sigma_{T-t_k}\lambda_{T-t_k}^{-1/2}\varepsilon_{\mathsf{score},k}^{1/2}\Big\}.$$

Per the discussions above, we see that $\mathbb{P}(\mathcal{S}^c_{Y_{t_k}, Y_{t_{k+1}}}) \le p_k(Y_{t_k}, Y_{t_{k+1}})$. For notational simplicity, we define

$$\begin{aligned}
N_k &:= \int \mathbb{1}\{g \in \mathcal{S}_{Y_{t_k}, Y_{t_{k+1}}}\} \exp\big(\eta_k\langle m(t_k, Y_{t_k} + \zeta_{k,1}g), v_k - \kappa_k g\rangle - \gamma_k\|m(t_k, Y_{t_k} + \zeta_{k,1}g)\|_2^2/2\big)\phi_{\xi_k,\tau_k}(g)\mathrm{d}g, \\
\widehat{N}_k &:= \int \mathbb{1}\{g \in \mathcal{S}_{Y_{t_k}, Y_{t_{k+1}}}\} \exp\big(\eta_k\langle \widehat{m}(t_k, Y_{t_k} + \zeta_{k,1}g), v_k - \kappa_k g\rangle - \gamma_k\|\widehat{m}(t_k, Y_{t_k} + \zeta_{k,1}g)\|_2^2/2\big)\phi_{\xi_k,\tau_k}(g)\mathrm{d}g, \\
N_k^c &:= \int \mathbb{1}\{g \in \mathcal{S}^c_{Y_{t_k}, Y_{t_{k+1}}}\} \exp\big(\eta_k\langle m(t_k, Y_{t_k} + \zeta_{k,1}g), v_k - \kappa_k g\rangle - \gamma_k\|m(t_k, Y_{t_k} + \zeta_{k,1}g)\|_2^2/2\big)\phi_{\xi_k,\tau_k}(g)\mathrm{d}g, \\
\widehat{N}_k^c &:= \int \mathbb{1}\{g \in \mathcal{S}^c_{Y_{t_k}, Y_{t_{k+1}}}\} \exp\big(\eta_k\langle \widehat{m}(t_k, Y_{t_k} + \zeta_{k,1}g), v_k - \kappa_k g\rangle - \gamma_k\|\widehat{m}(t_k, Y_{t_k} + \zeta_{k,1}g)\|_2^2/2\big)\phi_{\xi_k,\tau_k}(g)\mathrm{d}g, \\
\widehat{D}_k &:= \int \exp\big(\eta_k\langle \widehat{m}(t_k, Y_{t_k} + \zeta_{k,1}g), v_k - \kappa_k g\rangle - \gamma_k\|\widehat{m}(t_k, Y_{t_k} + \zeta_{k,1}g)\|_2^2/2\big)\phi_{\xi_k,\tau_k}(g)\mathrm{d}g,
\end{aligned}$$

which clearly obey $\widehat{D}_k^{-1}(\widehat{N}_k+\widehat{N}_k^c)=1$, and

$$\log\frac{\int\exp\big(\eta_k\langle m(t_k,Y_{t_k}+\zeta_{k,1}g),v_k-\kappa_k g\rangle-\gamma_k\|m(t_k,Y_{t_k}+\zeta_{k,1}g)\|_2^2/2\big)\phi_{\xi_k,\tau_k}(g)\mathrm{d}g}{\int\exp\big(\eta_k\langle \widehat{m}(t_k,Y_{t_k}+\zeta_{k,1}g),v_k-\kappa_k g\rangle-\gamma_k\|\widehat{m}(t_k,Y_{t_k}+\zeta_{k,1}g)\|_2^2/2\big)\phi_{\xi_k,\tau_k}(g)\mathrm{d}g}=\log\frac{N_k+N_k^c}{\widehat{D}_k}.$$

Hence, in order to upper bound term (ii) (cf. Eq. (61)), it suffices to upper bound $\widehat{D}_k^{-1}(N_k+N_k^c)$, towards which we intend to upper bound $\widehat{D}_k^{-1}|N_k-\widehat{N}_k|$ and $\widehat{D}_k^{-1}|N_k^c-\widehat{N}_k^c|$ separately.

Let us start with the first term $\widehat{D}_k^{-1}|N_k-\widehat{N}_k|$. Note that for $a_1,a_2\in\mathbb{R}$, we have $|e^a-e^b|\le e^{\max\{a,b\}}|a-b|$. As a consequence, when g falls inside $\mathcal{S}_{Y_{t_k},t_{k+1}}$, we have

$$\begin{aligned}&\widehat{D}_k^{-1}|N_k-\widehat{N}_k| \qquad (69)\\ &\le\frac{\varepsilon_{\mathsf{score},k}^{1/2}\sigma_{T-t_k}}{\widehat{D}_k\lambda_{T-t_k}^{1/2}}\int e^{2\eta_k\sqrt{d}(\|v_k\|_2+\kappa_k\|g-\xi_k\|_2+\kappa_k\|\xi_k\|_2)}(\eta_k\|v_k\|_2+\eta_k\|\kappa_k\xi_k\|_2+\eta_k\kappa_k\|g-\xi_k\|_2+3\gamma_k\sqrt{d}/2)\phi_{\xi_k,\tau_k}(g)\mathrm{d}g.\end{aligned}$$

Note that

$$\begin{aligned}&\int e^{2\eta_k\sqrt{d}(\|v_k\|_2+\kappa_k\|g-\xi_k\|_2+\kappa_k\|\xi_k\|_2)}(\eta_k\|v_k\|_2+\eta_k\|\kappa_k\xi_k\|_2+\eta_k\kappa_k\|g-\xi_k\|_2+3\gamma_k\sqrt{d}/2)\phi_{\xi_k,\tau_k}(g)\mathrm{d}g\\ &\le\int e^{2\eta_k\sqrt{d}(\|v_k\|_2+\kappa_k\|\xi_k\|_2)+4\eta_k\kappa_k d+\eta_k\kappa_k\|g-\xi_k\|_2^2/4}(\eta_k\|v_k\|_2+\eta_k\|\kappa_k\xi_k\|_2+\eta_k\kappa_k\|g-\xi_k\|_2+3\gamma_k\sqrt{d}/2)\phi_{\xi_k,\tau_k}(g)\mathrm{d}g\\ &\le\frac{e^{2\eta_k\sqrt{d}(\|v_k\|_2+\kappa_k\|\xi_k\|_2)+4\eta_k\kappa_k d}}{(1-\tau_k^2\eta_k\kappa_k/2)^{d/2}}\Big(\eta_k\|v_k\|_2+\eta_k\|\kappa_k\xi_k\|_2+3\gamma_k\sqrt{d}/2+\eta_k\kappa_k\big(\frac{d\tau_k^2}{1-\tau_k^2\eta_k\kappa_k/2}\big)^{1/2}\Big). \qquad (70)\end{aligned}$$

Note that the validity of Eq. (70) is conditional on $\eta_k\kappa_k\tau_k^2<2$: Only under this condition can we apply Gaussian integral to derive the last upper bound. We verify this condition in the next lemma.

Lemma C.4. *Under the assumptions of Lemma 4.4, it holds that* $1-\eta_k\kappa_k\tau_k^2/2\ge 0.4$ *for all* $0\le k\le K-1$.

Proof of Lemma C.4. Note that

$$\eta_k\kappa_k\tau_k^2=\frac{\kappa_k\lambda_{T-t_k}(e^{\Delta_k}-e^{-\Delta_k})}{\sigma_{T-t_k}^2(\zeta_{k,3}^2+\kappa_k^2)}.$$

Inspecting the proof of Lemma C.2, we see that: under the current assumptions, we have $\sigma_{T-t_k}^{-2}(e^{\Delta_k}-e^{-\Delta_k})\le 3.25\sqrt{\Delta_k\kappa}$, $\zeta_{k,1}\in[0.8\Delta_k^{1/2},0.9\Delta_k^{1/2}]$ and $\zeta_{k,3}\in[0.7\Delta_k^{1/2},0.75\Delta_k^{1/2}]$. As a consequence, one has

$$\Big|\frac{\kappa_k\lambda_{T-t_k}(e^{\Delta_k}-e^{-\Delta_k})}{\sigma_{T-t_k}^2(\zeta_{k,3}^2+\kappa_k^2)}\Big|\le\Big|\frac{(e^{\Delta_k}-e^{-\Delta_k})}{2\sigma_{T-t_k}^2\zeta_{k,3}}\Big|\le 2.4\sqrt{\kappa},$$

which is no larger than 1.2 when $\kappa<1/4$. □

Note that under Assumption C, the quantity $\widehat{D}_k$ admits the following lower bound:

$$\begin{aligned}\widehat{D}_k&\ge\int\exp\big(-2\eta_k\sqrt{d}\|v_k\|_2-2\eta_k\kappa_k\sqrt{d}\|g-\xi_k\|_2-2\eta_k\kappa_k\sqrt{d}\|\xi_k\|_2-2\gamma_k d\big)\phi_{\xi_k,\tau_k}(g)\mathrm{d}g\\ &\ge\exp\big(-2\eta_k\sqrt{d}\|v_k\|_2-\eta_k\kappa_k d-2\eta_k\kappa_k\sqrt{d}\|\xi_k\|_2-2\gamma_k d\big)/(1+\tau_k^2\eta_k\kappa_k)^d \qquad (71)\\ &\gtrsim\exp\big(-2\eta_k\sqrt{d}\|v_k\|_2-\eta_k\kappa_k d-2\eta_k\kappa_k\sqrt{d}\|\xi_k\|_2-2\gamma_k d\big),\end{aligned}$$

where the last inequality is due to the upper bound $(1+\tau_k^2\eta_k\kappa_k)^d\le\exp(d\tau_k^2\eta_k\kappa_k)$, which by the proof of Lemma C.4 is no larger than $\exp(2.4\sqrt{\kappa}d)\lesssim 1$. Here, we utilize the assumption that $\kappa d^2\lesssim 1$. Again by inspecting the proof of Lemma C.4, we see that $(1-\tau_k^2\eta_k\kappa_k/2)^{d/2}\ge(1-1.2\sqrt{\kappa})^{d/2}\ge\exp(-\sqrt{\kappa}d)$, which by

Assumption B is lower bounded by a positive numerical constant. Using this lower bound, we arrive at the following conclusion:

$$\begin{aligned}&\text{The last line of Eq. (70)}\\&\lesssim e^{2\eta_k\sqrt{d}(\|v_k\|_2+\kappa_k\|\xi_k\|_2)+4\eta_k\kappa_k d}\Big(\eta_k\|v_k\|_2+\eta_k\|\kappa_k\xi_k\|_2+3\gamma_k\sqrt{d}/2+\eta_k\kappa_k\tau_k d^{1/2}\Big).\end{aligned}\tag{72}$$

Putting together Eqs. (70) to (72) and using Eq. (67), we have

$$\begin{aligned}&\widehat{D}_k^{-1}|N_k-\widehat{N}_k|\\&\lesssim\frac{\varepsilon_{\mathsf{score},k}^{1/2}\sigma_{T-t_k}}{\lambda_{T-t_k}^{1/2}}e^{4\eta_k\sqrt{d}(\|v_k\|_2+\kappa_k\|\xi_k\|_2)+5\eta_k\kappa_k d+2\gamma_k d}\Big(\eta_k\|v_k\|_2+\eta_k\|\kappa_k\xi_k\|_2+3\gamma_k\sqrt{d}/2+\eta_k\kappa_k\tau_k d^{1/2}\Big).\end{aligned}$$

Taking the expectation over Q_{k+1} leads to

$$\mathbb{E}_{Q_{k+1}}\big[\widehat{D}_k^{-1}|N_k-\widehat{N}_k|\big]\lesssim\frac{\varepsilon_{\mathsf{score},k}^{1/2}\sigma_{T-t_k}}{\lambda_{T-t_k}^{1/2}}\exp(d\sqrt{\kappa})\kappa^{1/2}d^{1/2}\lesssim\frac{\varepsilon_{\mathsf{score},k}^{1/2}\kappa^{1/2}\sigma_{T-t_k}d^{1/2}}{\lambda_{T-t_k}^{1/2}}.\tag{73}$$

We then move on to control $\mathbb{E}_{Q_{k+1}}\big[\widehat{D}_k^{-1}|N_k^c-\widehat{N}_k^c|\big]$. Once again using the fact that for any $a_1,a_2\in\mathbb{R}$, we have $|e^a-e^b|\le e^{\max\{a,b\}}|a-b|$. As a result,

$$\begin{aligned}&\widehat{D}_k^{-1}|N_k^c-\widehat{N}_k^c|\\&\le\widehat{D}_k^{-1}\int\mathbb{1}\{g\in\mathcal{S}^c_{Y_{t_k},Y_{t_{k+1}}}\}e^{2\eta_k\sqrt{d}(\|v_k\|_2+\kappa_k\|g-\xi_k\|_2+\kappa_k\|\xi_k\|_2)}\big(3\eta_k\sqrt{d}(\|v_k\|_2+\kappa_k\|g\|_2)+2\gamma_k d\big)\phi_{\xi_k,\tau_k}(g)\mathrm{d}g\\&\le\widehat{D}_k^{-1}\mathbb{P}\big(\mathcal{S}^c_{Y_{t_k},Y_{t_{k+1}}}\big)^{1/2}\Big(\int e^{4\eta_k\sqrt{d}(\|v_k\|_2+\kappa_k\|g-\xi_k\|_2+\kappa_k\|\xi_k\|_2)}\big(3\eta_k\sqrt{d}(\|v_k\|_2+\kappa_k\|g\|_2)+2\gamma_k d\big)^2\phi_{\xi_k,\tau_k}(g)\mathrm{d}g\Big)^{1/2},\end{aligned}$$

where the last inequality above arises from the Cauchy-Schwarz inequality. Recall that $\mathbb{P}(\mathcal{S}^c_{Y_{t_k},Y_{t_{k+1}}})\le p_k(Y_{t_k},Y_{t_{k+1}})$, which satisfies $\mathbb{E}[p_k(Y_{t_k},Y_{t_{k+1}})]\le\sigma_{T-t_k}^2\lambda_{T-t_k}^{-1}\varepsilon_{\mathsf{score},k}$. Taking the expectation over Q_{k+1} gives

$$\mathbb{E}_{Q_{k+1}}\big[\widehat{D}_k^{-1}|N_k^c-\widehat{N}_k^c|\big]\lesssim\frac{\varepsilon_{\mathsf{score},k}^{1/2}\kappa^{1/2}\sigma_{T-t_k}d}{\lambda_{T-t_k}^{1/2}}.\tag{74}$$

Finally, put together Eqs. (73) and (74) to demonstrate that

$$\begin{aligned}(ii)&=\mathbb{E}_{Q_{k+1}}\big[\log\big(\widehat{D}_k^{-1}(N_k+N_k^c)\big)\big]\\&\le\mathbb{E}_{Q_{k+1}}\big[\log\big(1+\widehat{D}_k^{-1}|N_k-\widehat{N}_k|+\widehat{D}_k^{-1}|N_k^c-\widehat{N}_k^c|\big)\big]\\&\le\mathbb{E}_{Q_{k+1}}\big[\widehat{D}_k^{-1}|N_k-\widehat{N}_k|\big]+\mathbb{E}_{Q_{k+1}}\big[\widehat{D}_k^{-1}|N_k^c-\widehat{N}_k^c|\big]\\&\lesssim\frac{\varepsilon_{\mathsf{score},k}^{1/2}\kappa^{1/2}\sigma_{T-t_k}d}{\lambda_{T-t_k}^{1/2}},\end{aligned}$$

thus concluding the proof.

C.5 Proof of Lemma C.1

Proof of the first point

To prove the first point, we note that by Eq. (33),

$$\mathbb{T}_1=\sum_{k=0}^{K-1}\int_{t_k}^{t_{k+1}}(e^{t-t_k}-e^{-t+t_k})^2\mathbb{E}\big[\|s(t_k,Y_{t_k})\|_2^2\big]\mathrm{d}t$$

$$\lesssim \sum_{k=0}^{K-1} \Delta_k^3 \sigma_{T-t_k}^{-2} \mathbb{E}\big[\|\mathbb{E}[g \mid \lambda_{T-t_k}\theta + \sigma_{T-t_k} g = Y_{t_k}]\|_2^2\big]$$
$$\lesssim \sum_{k=0}^{K-1} \sigma_{T-t_k}^{-2} \Delta_k^3 d,$$

where the last upper bound is by Jensen's inequality.

Proof of the second point

By the triangle inequality, we can show that

$$\mathbb{T}_2 \le \underbrace{\sum_{k=0}^{K-1} \int_{t_k}^{t_{k+1}} \mathbb{E}\big[\big\|s(t, Y_t) - s(t_k, Y_{t_k}) - \sqrt{2}\nabla_x s(t_k, Y_{t_k})(B_t - B_{t_k})\big\|_2^2\big]\mathrm{d}t}_{(i)}$$
$$+ \underbrace{\sum_{k=0}^{K-1} \int_{t_k}^{t_{k+1}} \mathbb{E}\big[\big\|\nabla_x s(t_k, Y_{t_k} + b_{t_k,t})(B_t - B_{t_k}) - \nabla_x s(t_k, Y_{t_k})(B_t - B_{t_k})\big\|_2^2\big]\mathrm{d}t}_{(ii)}.$$

In what follows, let us upper bound terms (i) and (ii) separately.

To upper bound term (ii), we note that by the fundamental theorem of calculus, we have

$$\nabla_x s(t_k, Y_{t_k} + b_{t_k,t})(B_t - B_{t_k}) - \nabla_x s(t_k, Y_{t_k})(B_t - B_{t_k})$$
$$= \int_0^1 \nabla_x^2 s(t_k, Y_{t_k} + \eta b_{t_k,t})[b_{t_k,t} \otimes (B_t - B_{t_k})]\mathrm{d}\eta.$$

In view of the above decomposition, it suffices to separately upper bound $\mathbb{E}[\|\nabla_x^2 s(t_k, Y_{t_k})[b_{t_k,t} \otimes (B_t - B_{t_k})]\|_2^2]$ and $\mathbb{E}[\|(\nabla_x^2 s(t_k, Y_{t_k} + \eta b_{t_k,t}) - \nabla_x^2 s(t_k, Y_{t_k}))[b_{t_k,t} \otimes (B_t - B_{t_k})]\|_2^2]$. Let us start with the first term. Invoking the second point of Lemma B.2, we see that

$$\nabla_x^2 s(t_k, Y_{t_k}) = -\frac{1}{\sigma_{T-t_k}^3} \mathbb{E}\big[(g - \mathbb{E}[g \mid \lambda_{T-t_k}\theta + \sigma_{T-t_k} g = Y_{t_k}])^{\otimes 3} \mid \lambda_{T-t_k}\theta + \sigma_{T-t_k} g = Y_{t_k}\big], \tag{75}$$

where the expectation is over $(\theta, g) \sim q_0 \otimes \mathcal{N}(0, I_d)$. For simplicity, we write $z_1 = b_{t_k,t}$ and $z_2 = B_t - B_{t_k}$. Observe that z_1 and z_2 are jointly normal with mean zero, with $(z_1, z_2) \perp Y_{t_k}$. As for the covariance structure, by Lemma A.1, it holds that

$$\mathrm{Cov}[z_1, z_2] = 2\sqrt{2}(e^{t-t_k} - e^{-t+t_k})^{-1}(e^{t-t_k} - t + t_k - 1)I_d,$$
$$\mathrm{Cov}[z_1] = 8(e^{t-t_k} - e^{-t+t_k})^{-2} f_1(t - t_k) I_d,$$
$$\mathrm{Cov}[z_2] = (t - t_k) I_d.$$

For $i \in [d]$, we define $L_i = g_i - \mathbb{E}[g_i \mid \lambda_{T-t_k}\theta + \sigma_{T-t_k} g = Y_{t_k}]$. For all indices $j_1, j_2, \ell_1, \ell_2 \in [d]$ that are not paired up[1], it holds that

$$\mathbb{E}\big[\mathbb{E}\left[L_i L_{j_1} L_{\ell_1} \mid \lambda_{T-t_k}\theta + \sigma_{T-t_k} g = Y_{t_k}\right]\mathbb{E}\big[L_i L_{j_2} L_{\ell_2} \mid \lambda_{T-t_k}\theta + \sigma_{T-t_k} g = Y_{t_k}\big]\, z_{1,j_1} z_{2,\ell_1} z_{1,j_2} z_{2,\ell_2}\big] = 0.$$

Substituting the above equation into Eq. (75), and applying the Cauchy-Schwartz inequality and the Jensen inequality, we arrive at

$$\mathbb{E}\big[\|\nabla_x^2 s(t_k, Y_{t_k})[b_{t_k,t} \otimes (B_t - B_{t_k})]\|_{\mathrm{F}}^2\big] \lesssim \sigma_{T-t_k}^{-6} \Delta_k^2 d^3. \tag{76}$$

[1] If j_1, j_2, ℓ_1, ℓ_2 are paired up, then we must have $\{j_1, j_2, \ell_1, \ell_2\} = \{x, x, y, y\}$ or $\{x, x, x, x\}$ for some $x, y \in [d]$.

We then upper bound $\mathbb{E}[\|(\nabla_x^2 s(t_k, Y_{t_k} + \eta b_{t_k,t}) - \nabla_x^2 s(t_k, Y_{t_k}))[b_{t_k,t} \otimes (B_t - B_{t_k})]\|_2^2]$. Once again by the fundamental theorem of calculus, we have

$$\begin{aligned}
&\mathbb{E}[\|(\nabla_x^2 s(t_k, Y_{t_k} + \eta b_{t_k,t}) - \nabla_x^2 s(t_k, Y_{t_k}))[b_{t_k,t} \otimes (B_t - B_{t_k})]\|_2^2] \\
&\quad = \mathbb{E}\big[\| \int_0^1 \nabla_x^3 s(t_k, Y_{t_k} + \kappa\eta b_{t_k,t})[\eta b_{t_k,t} \otimes b_{t_k,t} \otimes (B_t - B_{t_k})]\mathrm{d}\kappa\|_2^2\big] \\
&\quad \lesssim \Delta_k^3 d^3 \int_0^1 \mathbb{E}[\|\nabla_x^3 s(t_k, Y_{t_k} + \kappa\eta b_{t_k,t})\|_{\mathrm{F}}^2]\mathrm{d}\kappa \\
&\quad \lesssim \lambda_{T-t_k}^8 \sigma_{T-t_k}^{-16} \Delta_k^3 d^7,
\end{aligned} \tag{77}$$

where the last inequality arises from Lemma B.5. Taking Eqs. (76) and (77) together, we derive an upper bound on term (ii) as follows:

$$(ii) \lesssim \sum_{k=0}^{K-1} \Big(\sigma_{T-t_k}^{-6} \Delta_k^3 d^3 + \lambda_{T-t_k}^8 \sigma_{T-t_k}^{-16} \Delta_k^4 d^7\Big). \tag{78}$$

We now turn attention to term (i). Define

$$E_{t_k,t} = \mathbb{E}\big[\|s(t, Y_t) - s(t_k, Y_{t_k}) - \sqrt{2}\nabla_x s(t_k, Y_{t_k})(B_t - B_{t_k})\|_2^2\big].$$

By the Itô formula, one has

$$\begin{aligned}
&s(t, Y_t) - s(t_k, Y_{t_k}) - \sqrt{2}\nabla_x s(t_k, Y_{t_k})(B_t - B_{t_k}) \\
&= \int_{t_k}^t \big[\partial_\tau s(\tau, Y_\tau) + \nabla_x s(\tau, Y_\tau)(Y_\tau + 2s(\tau, Y_\tau)) + \nabla_x^2 s(\tau, Y_\tau)[I_d]\big]\,\mathrm{d}\tau \\
&\quad + \sqrt{2}\int_{t_k}^t \big[\nabla_x s(\tau, Y_\tau) - \nabla_x s(t_k, Y_{t_k})\big]\mathrm{d}B_\tau.
\end{aligned}$$

Hence, in order to upper bound $E_{t_k,t}$, it suffices to control the following two quantities:

$$c_{t_k,t}^{(1)} = \mathbb{E}\left[\Big\| \int_{t_k}^t [\nabla_x s(\tau, Y_\tau) - \nabla_x s(t_k, Y_{t_k})]\mathrm{d}B_\tau\Big\|_2^2\right], \tag{80a}$$

$$c_{t_k,t}^{(2)} = \mathbb{E}\left[\Big\| \int_{t_k}^t \big[\partial_\tau s(\tau, Y_\tau) + \nabla_x s(\tau, Y_\tau)(Y_\tau + 2s(\tau, Y_\tau)) + \nabla_x^2 s(\tau, Y_\tau)[I_d]\big]\,\mathrm{d}\tau\Big\|_2^2\right], \tag{80b}$$

and then invoke $E_{t_k,t} \lesssim c_{t_k,t}^{(1)} + c_{t_k,t}^{(2)}$. In what follows, we upper bound $c_{t_k,t}^{(1)}$ and $c_{t_k,t}^{(2)}$ separately.

- Note that $c_{t_k,t_k}^{(1)} = 0$. Hence, in order to upper bound $c_{t_k,t}^{(1)}$, we can take the differential of $c_{t_k,t}^{(1)}$ with respect to t. Specifically, according to the Itô formula,

$$\begin{aligned}
\mathrm{d}c_{t_k,t}^{(1)} &= 2\mathbb{E}\left[\Big\langle \int_{t_k}^t [\nabla_x s(\tau, Y_\tau) - \nabla_x s(t_k, Y_{t_k})]\mathrm{d}B_\tau,\, [\nabla_x s(t, Y_t) - \nabla_x s(t_k, Y_{t_k})]\mathrm{d}B_t\Big\rangle\right] \\
&\quad + \mathbb{E}\left[\|\nabla_x s(t, Y_t) - \nabla_x s(t_k, Y_{t_k})\|_{\mathrm{F}}^2\right]\mathrm{d}t \\
&= 3\mathbb{E}\left[\|\nabla_x s(t, Y_t) - \nabla_x s(t_k, Y_{t_k})\|_{\mathrm{F}}^2\right]\mathrm{d}t.
\end{aligned} \tag{81}$$

  The term $\mathbb{E}[\|\nabla_x s(t, Y_t) - \nabla_x s(t_k, Y_{t_k})\|_{\mathrm{F}}^2]$ shall be bounded in the next lemma, whose proof is postponed to Appendix C.6.

  Lemma C.5. *For $t_k \le t \le t_{k+1}$, we define*

$$M_{t_k,t} = \mathbb{E}\left[\|\nabla_x s(t, Y_t) - \nabla_x s(t_k, Y_{t_k})\|_{\mathrm{F}}^2\right].$$

  Then, under the conditions of Lemma 4.3, for all $t \in [t_k, t_{k+1}]$ we have $M_{t_k,t} \lesssim d^3 \sigma_{T-t_{k+1}}^{-6} \Delta_k$.

Armed with Lemma C.5 and Eq. (81), we conclude that for all $t \in [t_k, t_{k+1}]$,

$$c_{t_k,t}^{(1)} \lesssim d^3 \sigma_{T-t_{k+1}}^{-6} \Delta_k^2. \tag{82}$$

- We now turn to establishing an upper bound on $c_{t_k,t}^{(2)}$, as defined in Eq. (80b). Leveraging the Cauchy–Schwarz inequality, we obtain that for all $t \in [t_k, t_{k+1}]$,

$$\begin{aligned} c_{t_k,t}^{(2)} &\lesssim \mathbb{E}\left[(t-t_k)\int_{t_k}^{t} \left\|\partial_\tau s(\tau, Y_\tau) + \nabla_x s(\tau, Y_\tau)\big(Y_\tau + 2s(\tau, Y_\tau)\big) + \nabla_x^2 s(\tau, Y_\tau)[I_d]\right\|_2^2 \mathrm{d}\tau\right] \\ &\lesssim (t-t_k)\int_{t_k}^{t} \Big(\mathbb{E}\big[\big\|\partial_\tau s(\tau, Y_\tau)\big\|_2^2\big] + \mathbb{E}\big[\big\|\nabla_x s(\tau, Y_\tau)\big(Y_\tau + s(\tau, Y_\tau)\big)\big\|_2^2\big] + \mathbb{E}\big[\big\|\nabla_x^2 s(\tau, Y_\tau)[I_d]\big\|_2^2\big]\Big)\, \mathrm{d}\tau. \end{aligned} \tag{83}$$

We develop separate upper bounds for the above summands in the lemma below; the proof is deferred to Appendix C.7.

Lemma C.6. *Under the conditions of Lemma 4.3, the following upper bounds hold for all $\tau \in [0, T)$:*

1. $\mathbb{E}\big[\|\partial_\tau s(\tau, Y_\tau)\|_2^2\big] \lesssim d^3 \lambda_{T-\tau}^4 \sigma_{T-\tau}^{-6} + d\lambda_{T-\tau}^2 \sigma_{T-\tau}^{-4}$;
2. $\mathbb{E}\big[\|\nabla_x s(\tau, Y_\tau) s(\tau, Y_\tau)\|_2^2\big] \lesssim d^3 \sigma_{T-\tau}^{-6}$;
3. $\mathbb{E}\big[\|\nabla_x s(\tau, Y_\tau) Y_\tau\|_2^2\big] \lesssim d^3 \sigma_{T-\tau}^{-4}$;
4. $\mathbb{E}\big[\|\nabla_x^2 s(\tau, Y_\tau)[I_d]\|_2^2\big] \lesssim \sigma_{T-\tau}^{-6} d^3$.

Substituting the upper bounds derived in Lemma C.6 into Eq. (83), we conclude that for all $t \in [t_k, t_{k+1}]$,

$$c_{t_k,t}^{(2)} \lesssim d^3 \sigma_{T-t_{k+1}}^{-6} \Delta_k^2. \tag{84}$$

Combining the preceding bounds in Eqs. (82) and (84), we obtain

$$E_{t_k,t} \lesssim c_{t_k,t}^{(1)} + c_{t_k,t}^{(2)} \lesssim d^3 \sigma_{T-t_{k+1}}^{-6} \Delta_k^2, \tag{85}$$

which further implies that

$$(i) \lesssim \sum_{k=0}^{K-1} d^3 \sigma_{T-t_{k+1}}^{-6} \Delta_k^3. \tag{86}$$

Putting Eqs. (78) and (86) together results in

$$\mathbb{T}_2 \lesssim \sum_{k=0}^{K-1} \Big(d^3 \sigma_{T-t_{k+1}}^{-6} \Delta_k^3 + d^7 \lambda_{T-t_k}^8 \sigma_{T-t_k}^{-16} \Delta_k^4\Big),$$

thus completing the proof.

Proof of the third point

From the triangle inequality, we obtain

$$\begin{aligned} \mathbb{T}_3 \lesssim &\sum_{k=0}^{K-1} \int_{t_k}^{t_{k+1}} \int_0^{t-t_k} \mathbb{E}\big[\big\|s(t_k + r, Y_{t_k+r}) - s(t_k, Y_{t_k}) - \sqrt{2}\nabla_x s(t_k, Y_{t_k})(B_{t_k+r} - B_{t_k})\big\|_2^2\big] \mathrm{d}t \\ &+ \sum_{k=0}^{K-1} \int_{t_k}^{t_{k+1}} \int_0^{t-t_k} \mathbb{E}\Big[\big\|\nabla_x s(t_k, Y_{t_k} + b_{t_k,t})(B_{t_k+r} - B_{t_k}) - \nabla_x s(t_k, Y_{t_k})(B_{t_k+r} - B_{t_k})\big\|_2^2\Big] \mathrm{d}t. \end{aligned}$$

Similar to the derivation of point 2, for all $t \in [0, t_{k+1} - t_k]$ and all $r \in [0, t - t_k]$ it holds that

$$\mathbb{E}\big[\big\|s(t_k + r, Y_{t_k+r}) - s(t_k, Y_{t_k}) - \sqrt{2}\nabla_x s(t_k, Y_{t_k})(B_{t_k+r} - B_{t_k})\big\|_2^2\big] \lesssim d^3\sigma_{T-t_{k+1}}^{-6}\Delta_k^2,$$
$$\mathbb{E}\big[\big\|\nabla_x s(t_k, Y_{t_k} + b_{t_k,t})(B_{t_k+r} - B_{t_k}) - \nabla_x s(t_k, Y_{t_k})(B_{t_k+r} - B_{t_k})\big\|_2^2\big] \lesssim \sigma_{T-t_k}^{-6}\Delta_k^2 d^3 + \lambda_{T-t_k}^8\sigma_{T-t_k}^{-16}\Delta_k^3 d^7.$$

The desired claim then follows.

Proof of the fourth point

This is similar to the proof of the second point. We skip the proof for the sake of brevity.

C.6 Proof of Lemma C.5

Proof of Lemma C.5. By Itô 's lemma we can write

$$\mathrm{d}\nabla_x s(t, Y_t) = \partial_t\nabla_x s(t, Y_t)\mathrm{d}t + \nabla_x^2 s(t, Y_t)(Y_t + 2s(t, Y_t))\mathrm{d}t + \sqrt{2}\nabla_x^2 s(t, Y_t)\mathrm{d}B_t + \nabla_x^3 s(t, Y_t)[I_d]\mathrm{d}t. \tag{87}$$

From the proof of Benton et al. (2024, Lemma 3), we deduce that for all $x \in \mathbb{R}^d$ and $t \in [0, T)$,

$$\partial_t s(t, x) = -\big[s(t, x) + \nabla_x s(t, x)x + \Delta s(t, x) + 2\nabla_x s(t, x)s(t, x)\big].$$

where Δ denotes the Laplace operator. Further taking the Jacobian of the above mapping, we obtain that

$$\partial_t\nabla_x s(t, x) = -\big[2\nabla_x s(t, x) + \nabla_x^2 s(t, x)x + \nabla_x^3 s(t, X_t)[I_d] + 2\nabla_x s(t, x)^2 + 2\nabla_x^2 s(t, x)s(t, x)\big]. \tag{88}$$

Putting together Eqs. (87) and (88), we derive

$$\mathrm{d}\nabla_x s(t, Y_t) = -\left[2\nabla_x s(t, Y_t) + 2\nabla_x s(t, Y_t)^2\right]\mathrm{d}t + \sqrt{2}\nabla_x^2 s(t, Y_t)\mathrm{d}B_t. \tag{89}$$

With Eq. (89), we can analyze the differential of M_{t,t_k} with respect to t. In particular,

$$\mathrm{d}M_{t_k,t} = -4\mathbb{E}\left[\left\langle\nabla_x s(t, Y_t) - \nabla_x s(t_k, Y_{t_k}), \nabla_x s(t, Y_t) + \nabla_x s(t, Y_t)^2\right\rangle\right]\mathrm{d}t + 2\mathbb{E}\left[\|\nabla_x^2 s(t, Y_t)\|_{\mathrm{F}}^2\right]\mathrm{d}t. \tag{90}$$

By Lemma B.2, we obtain that (recall $m_g(t, x) = \mathbb{E}[g \mid \lambda_{T-t}\theta + \sigma_{T-t}g = x]$)

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla_x^2 s(t, Y_t)\|_{\mathrm{F}}^2\right] = \frac{1}{\sigma_{T-t}^6}\mathbb{E}\left[\|\mathbb{E}[(g - m_g(t, Y_t))^{\otimes 3} \mid \lambda_{T-t}\theta + \sigma_{T-t}g = Y_t]\|_{\mathrm{F}}^2\right] \lesssim \frac{d^3}{\sigma_{T-t}^6},\\
&\mathbb{E}\left[\|\nabla_x s(t, Y_t)\|_{\mathrm{F}}^2\right] \lesssim \mathbb{E}\left[\|\sigma_{T-t}^{-2}I_d\|_{\mathrm{F}}^2\right] + \mathbb{E}\left[\|\sigma_{T-t}^{-2}\mathbb{E}[(g - m_g(t, Y_t))^{\otimes 2} \mid \lambda_{T-t}\theta + \sigma_{T-t}g = Y_t]\|_{\mathrm{F}}^2\right] \lesssim \frac{d^2}{\sigma_{T-t}^4},\\
&\mathbb{E}\left[\|\nabla_x s(t, Y_t)^2\|_{\mathrm{F}}^2\right] \lesssim \mathbb{E}\left[\|\sigma_{T-t}^{-4}I_d\|_{\mathrm{F}}^2\right] + \mathbb{E}\left[\|\sigma_{T-t}^{-4}\mathbb{E}[(g - m_g(t, Y_t))^{\otimes 2} \mid \lambda_{T-t}\theta + \sigma_{T-t}g = Y_t]^2\|_{\mathrm{F}}^2\right] \lesssim \frac{d^4}{\sigma_{T-t}^8}.
\end{aligned} \tag{91}$$

Applying the Cauchy-Schwartz inequality to Eq. (90) and substituting in the upper bounds from Eq. (91) give

$$\begin{aligned}
\mathrm{d}M_{t_k,t} &\lesssim \mathbb{E}\left[\|\nabla_x s(t, Y_t) - \nabla_x s(t_k, Y_{t_k})\|_{\mathrm{F}}^2\right]^{1/2} \cdot \mathbb{E}\left[\|\nabla_x s(t, Y_t) + \nabla_x s(t, Y_t)^2\|_{\mathrm{F}}^2\right]^{1/2}\mathrm{d}t + \mathbb{E}\left[\|\nabla_x^2 s(t, Y_t)\|_{\mathrm{F}}^2\right]\mathrm{d}t\\
&\lesssim \frac{d^3}{\sigma_{T-t}^6}\mathrm{d}t.
\end{aligned}$$

Observe that $M_{t_k,t_k} = 0$. As a consequence, for all $t \in [t_k, t_{k+1}]$, it holds that $M_{t_k,t} \lesssim d^3\sigma_{T-t_{k+1}}^{-6}\Delta_k$. The proof is complete. □

C.7 Proof of Lemma C.6

Proof of Lemma C.6, point 1. By the third point of Lemma B.2,

$$-\partial_\tau s(\tau, Y_\tau) = -\frac{\lambda_{T-\tau}}{\sigma_{T-\tau}^2}\mathbb{E}\big[\theta \mid \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau\big] + \frac{2\lambda_{T-\tau}^2}{\sigma_{T-\tau}^3}\mathbb{E}\big[g \mid \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau\big] + \frac{\lambda_{T-\tau}}{\sigma_{T-\tau}^2}\mathbb{E}\big[\big(\theta - m(\tau, Y_\tau)\big)\big(\mathcal{F}(\theta, Y_\tau, \tau) - m_{\mathcal{F}}\big) \mid \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau\big],$$

where $m_{\mathcal{F}} = \mathbb{E}[\mathcal{F}(\theta, Y_\tau, \tau) \mid \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau]$. Under Assumption A, by Jensen's inequality we have

$$\mathbb{E}\left[\|\mathbb{E}[\theta \mid \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau]\|_2^2\right] \lesssim d,$$
$$\mathbb{E}\left[\|\mathbb{E}[g \mid \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau]\|_2^2\right] \lesssim d.$$

Recall that $\mathcal{F}$ is defined in Eq. (34). Conditional on $\lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau$, we have

$$\begin{aligned}
&\mathcal{F}(\theta, Y_\tau, \tau) - \mathbb{E}\big[\mathcal{F}(\theta, Y_\tau, \tau) \mid \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau\big] \\
&= \frac{\lambda_{T-\tau}^2}{\sigma_{T-\tau}^4}\|\theta\|_2^2 - \frac{\lambda_{T-\tau} + \lambda_{T-\tau}^3}{\sigma_{T-\tau}^4}\langle Y_\tau, \theta\rangle \\
&\quad - \mathbb{E}\left[\frac{\lambda_{T-\tau}^2}{\sigma_{T-\tau}^4}\|\theta\|_2^2 - \frac{\lambda_{T-\tau} + \lambda_{T-\tau}^3}{\sigma_{T-\tau}^4}\langle Y_\tau, \theta\rangle \,\middle|\, \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau\right] \\
&= -\frac{\lambda_{T-\tau}}{\sigma_{T-\tau}}\langle\theta, g\rangle + \frac{\lambda_{T-\tau}^2}{\sigma_{T-\tau}^2}\|g\|_2^2 + \mathbb{E}\left[\frac{\lambda_{T-\tau}}{\sigma_{T-\tau}}\langle\theta, g\rangle - \frac{\lambda_{T-\tau}^2}{\sigma_{T-\tau}^2}\|g\|_2^2 \,\middle|\, \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau\right].
\end{aligned} \tag{92}$$

Note that conditioning on θ, $\langle\theta, g\rangle$ has conditional distribution $\mathcal{N}(0, \|\theta\|_2^2)$. Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbb{E}\big[\big(\theta - m(\tau, Y_\tau)\big)\big(\mathcal{F}(\theta, Y_\tau, \tau) - m_{\mathcal{F}}\big) \mid \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau\big]\right\|_2^2\right] \\
&\lesssim \mathbb{E}\left[\mathbb{E}\big[\|\theta - m(\tau, Y_\tau)\|_2^2 \mid \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau\big]\mathbb{E}\big[\big(\mathcal{F}(\theta, Y_\tau, \tau) - m_{\mathcal{F}}\big)^2 \mid \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau\big]\right] \\
&\lesssim \mathbb{E}\left[\left\|\Theta - m(\tau, Y_\tau)\right\|_2^4\right]^{1/2}\mathbb{E}\left[\big(\mathcal{F}(\Theta, Y_\tau, \tau) - m_{\mathcal{F}}\big)^4\right]^{1/2} \\
&\lesssim \frac{\lambda_{T-\tau}^2 d^3}{\sigma_{T-\tau}^2},
\end{aligned}$$

where we write $Y_\tau = \lambda_{T-\tau}\Theta + \sigma_{T-\tau}G$ for $(\Theta, G) \sim q_0 \otimes \mathcal{N}(0, 1)$. In the above display, the last inequality is by Jensen's inequality and Assumption A. The proof is thus complete. □

Proof of Lemma C.6, point 2. By Lemma B.2, we have

$$\begin{aligned}
&\mathbb{E}[\|\nabla_x s(\tau, Y_\tau)s(\tau, Y_\tau)\|_2^2] \\
&= \mathbb{E}\left[\|(\sigma_{T-\tau}^{-3}I_d - \lambda_{T-\tau}^2\sigma_{T-\tau}^{-5}\operatorname{Cov}[\theta \mid \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau])\mathbb{E}[g \mid \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau]\|_2^2\right] \\
&\lesssim \frac{1}{\sigma_{T-\tau}^6}\mathbb{E}\left[\|\mathbb{E}[g \mid \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau]\|_2^2\right] \\
&\quad + \frac{1}{\sigma_{T-\tau}^6}\mathbb{E}\left[\|\mathbb{E}[(g - m_g(\tau, Y_\tau))^{\otimes 2} \mid \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau]\mathbb{E}[g \mid \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau]\|_2^2\right] \\
&\lesssim d^3\sigma_{T-\tau}^{-6},
\end{aligned}$$

where we recall that $m_g(\tau, Y_\tau) = \mathbb{E}[g \mid \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau]$. □

Proof of Lemma C.6, point 3. Next, we look at the third term $\mathbb{E}[\|\nabla_x s(\tau, Y_\tau)Y_\tau\|_2^2]$. By the first point of Lemma B.2, we have

$$\mathbb{E}\left[\left\|\nabla_x s(\tau, Y_\tau)Y_\tau\right\|_2^2\right] = \mathbb{E}\left[\left\|\left(\sigma_{T-\tau}^{-2}I_d - \lambda_{T-\tau}^2\sigma_{T-\tau}^{-4}\operatorname{Cov}[\theta \mid \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau]\right)Y_\tau\right\|_2^2\right]$$

$$\lesssim \frac{1}{\sigma_{T-\tau}^4}\mathbb{E}[\|Y_\tau\|_2^2] + \frac{1}{\sigma_{T-\tau}^4}\mathbb{E}\left[\|\mathbb{E}[(g - m_g(\tau, Y_\tau))^{\otimes 2} \mid \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau]Y_\tau\|_2^2\right]$$
$$\leq \frac{1}{\sigma_{T-\tau}^4}\mathbb{E}[\|Y_\tau\|_2^2] + \frac{1}{\sigma_{T-\tau}^4}\mathbb{E}\left[\|\mathbb{E}[gg^\top \mid \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau]Y_\tau\|_2^2\right]. \tag{93}$$

Given that $(\mathbb{E}[M \mid \mathcal{F}])^2 \preceq \mathbb{E}[M^2 \mid \mathcal{F}]$ for any random matrix M and filtration $\mathcal{F}$, we can derive

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbb{E}[gg^\top \mid \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau]Y_\tau\right\|_2^2\right] &\leq \mathbb{E}\big[Y_\tau^\top \mathbb{E}\big[\|g\|_2^2 gg^\top \mid \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau\big]Y_\tau\big] \\ &= \mathbb{E}\big[(Y_\tau^\top G)^2\|G\|_2^2\big],\end{aligned} \tag{94}$$

where $Y_\tau = \lambda_{T-\tau}\Theta + \sigma_{T-\tau}G$ for $(\Theta, G) \sim q_0 \otimes \mathcal{N}(0, I_d)$. Also, observe that

$$\mathbb{E}\big[(Y_\tau^\top G)^2\|G\|_2^2\big] \lesssim d^3. \tag{95}$$

Substituting Eqs. (94) and (95) into Eq. (93) gives

$$\mathbb{E}\big[\big\|\nabla_x s(\tau, Y_\tau)Y_\tau\big\|_2^2\big] \lesssim \frac{d^3}{\sigma_{T-\tau}^4},$$

which completes the proof of the lemma. □

Proof of Lemma C.6, point 4. Finally, we upper bound $\mathbb{E}[\|\nabla_x^2 s(\tau, Y_\tau)[I_d]\|_2^2]$. By the second point of Lemma B.2, the i-th entry of $\nabla_x^2 s(\tau, X_\tau)[I_d]$ admits the form

$$\sum_{j\in[d]} \sigma_{T-\tau}^{-3}\mathbb{E}[(g_i - m_g(\tau, Y_\tau)_i)(g_j - m_g(\tau, Y_\tau)_j)^2 \mid \lambda_{T-\tau}\theta + \sigma_{T-\tau}g = Y_\tau], \tag{96}$$

where we recall that $m_g(t, x) = \mathbb{E}[g \mid \lambda_{T-t}\theta + \sigma_{T-t}g = x]$, $m_g(t, x)_i$ is the i-th entry of $m_g(t, x)$ and g_i is the i-th entry of g. Applying Jensen's inequality to Eq. (96), we conclude that $\mathbb{E}[\|\nabla_x^2 s(\tau, Y_\tau)[I_d]\|_2^2] \lesssim \sigma_{T-\tau}^{-6} d^3$. This concludes the proof. □

C.8 Proof of Corollary 3.5

Given δ and κ, we first construct a sequence of step sizes as follows:

- set $\Delta_{K-1} = \kappa\delta^2$.
- For $k = K-1, K-2, \cdots, 1$,

$$\Delta_{k-1} = \begin{cases} \Delta_k(1 + \sqrt{\kappa\Delta_k})^2, & \text{if } \Delta_k(1 + \sqrt{\kappa\Delta_k})^2 \leq \kappa; \\ \kappa, & \text{else.} \end{cases} \tag{97}$$

Next, we prove that the step sizes defined as above satisfies

$$\Delta_k \leq \kappa \min\left\{1, (T - t_{k+1})^2\right\}, \qquad k = 0, 1, \cdots, K-1.$$

Let us prove this claim by induction. When $k = K-1$, this is true by definition. Now suppose $\Delta_k \leq \kappa\min\{1, (T-t_{k+1})^2\}$ for some k, and we shall use this upper bound to prove $\Delta_{k-1} \leq \kappa\min\{1, (T-t_k)^2\}$. If $\Delta_k(1 + \sqrt{\kappa\Delta_k})^2 > \kappa$, then this is automatically true as $\Delta_{k-1} = \kappa$. Otherwise if $\Delta_k(1 + \sqrt{\kappa\Delta_k})^2 \leq \kappa$, we have $\Delta_{k-1} = \Delta_k(1 + \sqrt{\kappa\Delta_k})^2$. By induction hypothesis, $T - t_{k+1} \geq \sqrt{\Delta_k/\kappa}$. Therefore,

$$T - t_k = T - t_{k+1} + \Delta_k \geq \sqrt{\Delta_k/\kappa} + \Delta_k = \sqrt{\Delta_{k-1}/\kappa}.$$

In this case, $\Delta_{k-1} \geq \Delta_k$ holds for all $k = 1, 2, \cdots, K-1$, which further implies that $\Delta_k \geq \kappa\delta^2$ for all $k = 0, 1, \cdots, K-1$. As a consequence, $\Delta_k \geq \Delta_{K-1}(1 + \kappa\delta)^{K-k-1} = \kappa\delta^2(1 + \kappa\delta)^{K-k-1}$.

We then upper bound the number of steps K needed as a function of δ, κ and T. Let $K_1 = \sup\{k : \Delta_k = \kappa\}$, then $K \le T/\kappa + K - K_1$. Recall that $\Delta_k \ge \kappa\delta^2(1+\kappa\delta)^{K-k-1}$, hence

$$K - K_1 \lesssim \frac{1}{\kappa\delta}\log(1+1/\delta).$$

We define

$$T = \frac{1}{2}\log(d/\varepsilon^2), \qquad \kappa = \min\left\{\frac{\varepsilon}{d^{3/2}T^{1/2}},\ \frac{1}{d^2}\right\}.$$

By definition, we have $\kappa d^2 \le 1$.

When $\varepsilon \le \sqrt{d}/2$, it holds that $T \gtrsim 1$. With T and κ selected as above, we have $d^3\kappa^2 T \le \varepsilon^2$ and $de^{-2T} = \varepsilon^2$. In addition, it is seen that

$$d^7\kappa^3\delta^{-1} \lesssim d^{5/2}\varepsilon^3\delta^{-1}, \qquad d^7\kappa^3 T \lesssim d^{5/2}\varepsilon^3.$$

Also, note that

$$\sum_{k=0}^{K-1}\frac{\varepsilon_{\mathsf{score},k}^{1/2}\kappa^{1/2}\sigma_{T-t_k}d}{\lambda_{T-t_k}^{1/2}} \le \sum_{k=0}^{K-1}\frac{\varepsilon_{\mathsf{score},k}^{1/2}\varepsilon^{1/2}de^{T/2}}{d^{3/4}T^{1/4}} \lesssim \sum_{k=0}^{K-1} d^{1/2}\varepsilon_{\mathsf{score},k}^{1/2}.$$

Combining the preceding upper bounds yields

$$\mathsf{KL}(q_\delta \parallel p_{\mathsf{output}}) \lesssim \varepsilon^2 + \varepsilon^3 d^{5/2}\delta^{-1} + \sum_{k=0}^{K-1} d^{1/2}\varepsilon_{\mathsf{score},k}^{1/2}.$$

In this case, $K \lesssim \frac{1}{\kappa\delta}\log(1+1/\delta) + \frac{1}{2\kappa}\log(d/\varepsilon^2)$.

- If $\varepsilon \le 1/\sqrt{d}$, then

$$K \lesssim \frac{d^{3/2}}{\delta\varepsilon}\log(1+1/\delta)\sqrt{\log(d/\varepsilon^2)} + \frac{d^{3/2}}{\varepsilon}\big[\log(d/\varepsilon^2)\big]^{3/2} = \widetilde{O}\big(d^{3/2}(\varepsilon\delta)^{-1}\big).$$

- In addition, if $\varepsilon \ge 1/\sqrt{d}$, then

$$K \lesssim \frac{d^2}{\delta}\log(1+1/\delta)\sqrt{\log(d/\varepsilon^2)} + d^2\big[\log(d/\varepsilon^2)\big]^{3/2} = \widetilde{O}\big(d^2\delta^{-1}\big).$$

The proof is thus complete.

References

Juan Miguel Lopez Alcaraz and Nils Strodthoff. Diffusion-based time series imputation and forecasting with structured state space models. *arXiv preprint arXiv:2208.09399*, 2022.

Namrata Anand and Tudor Achim. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv preprint arXiv:2205.15019*, 2022.

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

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34: 17981–17993, 2021.

Joe Benton, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods. *arXiv preprint arXiv:2305.16860*, 2023.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, 2024.

Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative modeling with denoising auto-encoders and langevin sampling. *arXiv preprint arXiv:2002.00107*, 2020.

Kevin Burrage and Pamela Marion Burrage. High strong order explicit runge-kutta methods for stochastic ordinary differential equations. *Applied Numerical Mathematics*, 22(1-3):81–101, 1996.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pages 4735–4763. PMLR, 2023a.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pages 4672–4712. PMLR, 2023b.

Minshuo Chen, Song Mei, Jianqing Fan, and Mengdi Wang. An overview of diffusion models: Applications, guided generation, statistical rates and optimization. *arXiv preprint arXiv:2404.07771*, 2024a.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023c.

Sitan Chen, Giannis Daras, and Alex Dimakis. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for ddim-type samplers. In *International Conference on Machine Learning*, pages 4462–4484. PMLR, 2023d.

Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ode is provably fast. *Advances in Neural Information Processing Systems*, 36, 2024b.

Sitan Chen, Vasilis Kontonis, and Kulin Shah. Learning general gaussian mixtures with efficient score matching. *arXiv preprint arXiv:2404.18893*, 2024c.

Muthu Chidambaram, Khashayar Gatmiry, Sitan Chen, Holden Lee, and Jianfeng Lu. What does guidance do? a fine-grained analysis in a simple setting. *arXiv preprint arXiv:2409.13074*, 2024.

Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Zehao Dou, Subhodh Kotekal, Zhehao Xu, and Harrison H Zhou. From optimal score matching to optimal sampling. *arXiv preprint arXiv:2409.07032*, 2024.

Oliver Y Feng, Yu-Chun Kao, Min Xu, and Richard J Samworth. Optimal convex m-estimation via score matching. *arXiv preprint arXiv:2403.16688*, 2024.

Xuefeng Gao and Lingjiong Zhu. Convergence analysis for general probability flow odes of diffusion models in wasserstein distances. *arXiv preprint arXiv:2401.17958*, 2024.

Martin Gonzalez, Nelson Fernandez Pinto, Thuy Tran, Hatem Hajri, Nader Masmoudi, et al. Seeds: Exponential sde solvers for fast high-quality sampling from diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Shivam Gupta, Linda Cai, and Sitan Chen. Faster diffusion-based sampling with randomized midpoints: Sequential and parallel. *arXiv preprint arXiv:2406.00924*, 2024.

Yinbin Han, Meisam Razaviyayn, and Renyuan Xu. Neural network-based score estimation in diffusion models: Optimization and generalization. *arXiv preprint arXiv:2401.15604*, 2024.

Philippe Hansen-Estruch, Ilya Kostrikov, Michael Janner, Jakub Grudzien Kuba, and Sergey Levine. Idql: Implicit q-learning as an actor-critic method with diffusion policies. *arXiv preprint arXiv:2304.10573*, 2023.

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

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23 (47):1–33, 2022.

Marlis Hochbruck and Alexander Ostermann. Exponential integrators. *Acta Numerica*, 19:209–286, 2010.

Daniel Zhengyu Huang, Jiaoyang Huang, and Zhengjiang Lin. Convergence analysis of probability flow ode for score-based generative models. *arXiv preprint arXiv:2404.09730*, 2024.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Alexia Jolicoeur-Martineau, Ke Li, Rémi Piché-Taillefer, Tal Kachman, and Ioannis Mitliagkas. Gotta go fast when generating data with score-based models. *arXiv preprint arXiv:2105.14080*, 2021.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Sungwon Kim, Heeseung Kim, and Sungroh Yoon. Guided-tts 2: A diffusion model for high-quality adaptive text-to-speech with untranscribed data. *arXiv preprint arXiv:2205.15370*, 2022.

Diederik P Kingma. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.

Frederic Koehler, Alexander Heckett, and Andrej Risteski. Statistical efficiency of score matching: The view from isoperimetry. *arXiv preprint arXiv:2210.00726*, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Wilhelm Kutta. *Beitrag zur näherungsweisen Integration totaler Differentialgleichungen*. Teubner, 1901.

Jean-François Le Gall. *Brownian motion, martingales, and stochastic calculus*. Springer, 2016.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882, 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985. PMLR, 2023.

Gen Li and Yuling Yan. Adapting to unknown low-dimensional structures in score-based diffusion models. *arXiv preprint arXiv:2405.14861*, 2024a.

Gen Li and Yuling Yan. $o(d/t)$ convergence theory for diffusion probabilistic models under minimal assumptions. *arXiv preprint arXiv:2409.18959*, 2024b.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards faster non-asymptotic convergence for diffusion-based generative models. *arXiv preprint arXiv:2306.09251*, 2023.

Gen Li, Yu Huang, Timofey Efimov, Yuting Wei, Yuejie Chi, and Yuxin Chen. Accelerating convergence of score-based diffusion models, provably. *arXiv preprint arXiv:2403.03852*, 2024a.

Gen Li, Zhihan Huang, and Yuting Wei. Towards a mathematical theory for consistency training in diffusion models. *arXiv preprint arXiv:2402.07802*, 2024b.

Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. A sharp convergence theory for the probability flow odes of diffusion models. *arXiv preprint arXiv:2408.02320*, 2024c.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022.

Yuchen Liang, Peizhong Ju, Yingbin Liang, and Ness Shroff. Non-asymptotic convergence of discrete-time diffusion models: New approach and improved rate. *arXiv preprint arXiv:2402.13901*, 2024.

Cheng Lu, Kaiwen Zheng, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Maximum likelihood training for score-based diffusion odes by high order denoising score matching. In *International Conference on Machine Learning*, pages 14429–14460. PMLR, 2022a.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022b.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022c.

Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021.

Song Mei and Yuchen Wu. Deep networks as denoising algorithms: Sample-efficient learning of diffusion models in high-dimensional graphical models. *arXiv preprint arXiv:2309.11420*, 2023.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pages 26517–26582. PMLR, 2023.

Bernt Øksendal. *Stochastic differential equations*. Springer, 2003.

Tim Pearce, Tabish Rashid, Anssi Kanervisto, Dave Bignell, Mingfei Sun, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Ida Momennejad, Katja Hofmann, et al. Imitating human behaviour with diffusion models. *arXiv preprint arXiv:2301.10677*, 2023.

Yury Polyanskiy. Information theory methods in statistics and computer science. *MIT course page*, 2020.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Carl Runge. Über die numerische auflösung von differentialgleichungen. *Mathematische Annalen*, 46(2): 167–178, 1895.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

Robin San-Roman, Eliya Nachmani, and Lior Wolf. Noise estimation for generative diffusion models. *arXiv preprint arXiv:2104.02600*, 2021.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

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*, 2020b.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 32211–32252, 2023.

Rong Tang and Yun Yang. Adaptivity of diffusion models to manifold structures. In *International Conference on Artificial Intelligence and Statistics*, pages 1648–1656. PMLR, 2024.

Wenpin Tang and Hanyang Zhao. Score-based diffusion models via stochastic differential equations–a technical tutorial. *arXiv preprint arXiv:2402.07487*, 2024.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems*, 34: 24804–24816, 2021.

Brian L Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi S Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. In *The Eleventh International Conference on Learning Representations*.

Andre Wibisono, Yihong Wu, and Kaylee Yingxi Yang. Optimal score estimation via empirical bayes smoothing. *arXiv preprint arXiv:2402.07747*, 2024.

Yuchen Wu, Minshuo Chen, Zihao Li, Mengdi Wang, and Yuting Wei. Theoretical insights for diffusion guidance: A case study for gaussian mixture models. *arXiv preprint arXiv:2403.01639*, 2024.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.

Shuchen Xue, Mingyang Yi, Weijian Luo, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhi-Ming Ma. Sa-solver: Stochastic adams solver for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

Kaihong Zhang, Heqi Yin, Feng Liang, and Jingbo Liu. Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. *arXiv preprint arXiv:2402.15602*, 2024.

Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. In *The Eleventh International Conference on Learning Representations*, 2023.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Kaiwen Zheng, Cheng Lu, Jianfei Chen, and Jun Zhu. Dpm-solver-v3: Improved diffusion ode solver with empirical model statistics. *Advances in Neural Information Processing Systems*, 36:55502–55542, 2023.