Low-dimensional adaptation of diffusion models: Convergence in total variation

Jiadong Liang* Zhihan Huang* Yuxin Chen*

January 22, 2025

Abstract

This paper investigates how diffusion generative models leverage (unknown) low-dimensional structure to accelerate sampling. Focusing on two mainstream samplers — the denoising diffusion implicit model (DDIM) and the denoising diffusion probabilistic model (DDPM) — and assuming accurate score estimates, we prove that their iteration complexities are no greater than the order of k/ε (up to some log factor), where ε is the precision in total variation distance and k is some intrinsic dimension of the target distribution. Our results are applicable to a broad family of target distributions without requiring smoothness or log-concavity assumptions. Further, we develop a lower bound that suggests the (near) necessity of the coefficients introduced by Ho et al. (2020) and Song et al. (2020) in facilitating low-dimensional adaptation. Our findings provide the first rigorous evidence for the adaptivity of the DDIM-type samplers to unknown low-dimensional structure, and improve over the state-of-the-art DDPM theory regarding total variation convergence.

Contents

1 Introduction **2**
- 1.1 Score-based generative modeling: DDPM and DDIM . . . 2
- 1.2 Harnessing low-dimensional structure? . . . 3
- 1.3 This paper . . . 4

2 Preliminaries **5**

3 Main results **8**
- 3.1 Key assumptions . . . 8
- 3.2 Convergence theory for DDIM . . . 9
- 3.3 Convergence theory for DDPM . . . 10
- 3.4 Interpretation from the lens of differential equations . . . 11
- 3.5 Other alternatives of coefficient design? . . . 13

4 Related work **14**

5 Discussion **15**

A Technical lemmas **20**

*Department of Statistics and Data Science, the Wharton School, University of Pennsylvania; email: {jdl97,zhihanh,yuxinc}@wharton.upenn.edu.

B Analysis for DDIM (proof of Theorem 1) **22**
B.1 Controlling the function $-W(x_t)$ . . . 27
B.2 Controlling the term related to score estimation errors . . . 27
B.3 Controlling the TV distance between X_{t-1} and $\Phi_t(X_t)$ . . . 30

C Analysis for DDPM (proof of Theorem 3) **33**
C.1 Introducing auxiliary sequences . . . 34
C.2 Analysis of $\mathcal{R}_{t-1}(x_{t-1})$ . . . 35
C.3 Determining the surrogate distribution . . . 36
C.4 Bounding the residual $\mathcal{G}(x_t, x_0) - 1$ . . . 37
C.5 Controlling the discretization error . . . 38
C.6 Controlling the term related to score estimation errors . . . 40

D Equivalence between relation (26) and Song et al. (2020, Eq. (12)) **41**

E Proofs about reverse-time differential equations **42**
E.1 Generalized reverse-time differential equations . . . 42
E.2 Proof of Proposition 2 . . . 43
E.3 Proof of Proposition 1 . . . 45

F Proof of the lower bound in Theorem 4 **45**

G Auxiliary lemmas and related proofs **46**
G.1 Proof of Lemma 1 . . . 46
G.2 Proof of Lemma 2 . . . 48
G.3 Proof of Lemma 3 . . . 49
G.4 Proof of Lemma 4 . . . 51
G.5 Proof of Lemma 5 . . . 52

1 Introduction

As a cornerstone of the rapidly evolving field of generative AI, diffusion generative models have driven mind-blowing progress across a diverse range of applications, such as image and video generation, medical image analysis, and time-series forecasting, to name just a few (Ramesh et al., 2022; Croitoru et al., 2023; Kazerouni et al., 2023; Lin et al., 2024; Yang et al., 2023). The remarkable effectiveness of diffusion models has inspired a recent wave of activity aimed at developing and strengthening their theoretical underpinnings.

1.1 Score-based generative modeling: DDPM and DDIM

At their core, diffusion models seek to gradually transform pure noise into new samples that emulate a d-dimensional target distribution p_{data}, accomplished by learning to reverse a forward stochastic process that progressively converts data into noise, detailed below.

Forward process. A common choice of the forward process with finite horizon T is given by

$$X_0 \sim p_{\mathsf{data}}, \qquad X_t = \sqrt{\alpha_t}\, X_{t-1} + \sqrt{1-\alpha_t}\, W_t, \quad t = 1, \cdots, T, \tag{1}$$

where $\{W_t\}_{t=1}^T$ comprises independent noise vectors obeying $W_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$, and the sequence $\{\alpha_t\}_{t=1}^T \subseteq (0,1)$ controls the variance of the Gaussian noise injected in each step. Informally, as T grows, the distribution of X_t typically converges rapidly to the standard Gaussian $\mathcal{N}(0, I_d)$.

Reverse process and diffusion-based samplers. As it turns out, the forward Markov process (1) is reversible in general, a property that follows from classical results in the stochastic differential equation (SDE) literature (Anderson, 1982; Haussmann and Pardoux, 1986). This intriguing property underpins the data generation process of diffusion models, which involves constructing a reverse process $Y_T \to \cdots \to Y_1 \to Y_0$ that closely mimics the forward process (1) in the sense that $Y_t \overset{\mathrm{d}}{\approx} X_t$ for each step t. Crucially, the reversal of the forward process hinges upon access to the so-called (Stein) score function

$$s_t^*(X) \coloneqq \nabla \log p_{X_t}(X) \tag{2}$$

— hence the term "score-based generative modeling." To formalize the sampling process, one needs to specify the initialization and iterative steps of the reverse process. The initialization step is straightforward: given that X_T is approximately Gaussian for large enough T, one generic choice is to draw Y_T as pure noise $\mathcal{N}(0, I_d)$. As such, a key step underlying the design of the sampling process boils down to how to update Y_t at each step while maintaining the desired distributional proximity. In what follows, we single out two mainstream paradigms, assuming availability of an estimate s_t of the true score function s_t^* at each t:

- *Denoising Diffusion Implicit Model (DDIM).* The DDIM sampler (or the probability flow ODE sampler) (Song et al., 2020) adopts a deterministic update rule below:

$$Y_T \sim \mathcal{N}(0, I_d), \qquad Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\big(Y_t + \eta_t^{\mathsf{ddim}} s_t(Y_t)\big), \qquad t = T, \cdots, 1, \tag{3}$$

  where $\{\eta_t^{\mathsf{ddim}}\}$ represents some suitably chosen coefficients. In words, each step (3) computes Y_{t-1} as a weighted sum of Y_t and its score estimate.

- *Denoising Diffusion Probabilistic Model (DDPM).* Originally proposed by Ho et al. (2020) as a way to optimize certain variational lower bounds on the log-likelihood, DDPM employs the following stochastic iterative updates:

$$Y_T \sim \mathcal{N}(0, I_d), \qquad Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\big(Y_t + \eta_t^{\mathsf{ddpm}} s_t(Y_t) + \sigma_t^{\mathsf{ddpm}} Z_t\big), \quad t = T, \cdots, 1, \tag{4}$$

  where the Z_t's are independently generated obeying $Z_t \sim \mathcal{N}(0, I_d)$, and $\{\eta_t^{\mathsf{ddpm}}\}$ and $\{\sigma_t^{\mathsf{ddpm}}\}$ are properly chosen coefficients. A key distinction from DDIM is that the iterative updates (4) inject additional stochastic noise at each step.

1.2 Harnessing low-dimensional structure?

Motivated by the practical efficacy of diffusion models, the past few years have witnessed a flurry of activity towards establishing convergence theory for both DDPM and DDIM (Lee et al., 2022; Chen et al., 2022b, 2023a,d, 2024b; Benton et al., 2024, 2023; Li et al., 2024b,c,d; Gao and Zhu, 2024; Huang et al., 2024a; Li and Yan, 2024b; Tang, 2023; Tang and Zhao, 2024b; Liang et al., 2024; Li and Jiao, 2024). For a fairly general family of target distributions p_{data} (without assuming smoothness and log-concavity), the state-of-the-art theory Li and Yan (2024a); Li et al. (2024c) demonstrated that for both DDPM and DDIM, it takes at most the order of (modulo some log factor)

$$\frac{d}{\varepsilon} \text{ iterations} \tag{5}$$

to yield a sample whose distribution is ε-close in total variation (TV) distance to the target distribution, provided that perfect score function estimates are available.

Nevertheless, even linear scaling in the ambient dimension d can still be prohibitively expensive for many contemporary applications. Take the ImageNet dataset (Deng et al., 2009) for instance: each image might contain 150,528 pixels, while its intrinsic dimension is estimated to be 43 or less (Pope et al., 2021). As

a result, applying the state-of-the-art theory (5) could suggest an iteration complexity that exceeds one million, even though practical implementations of DDIM and DDPM often produce high-quality samples in just a few hundred (or even a few ten) iterations. The discrepancy between theory and practice suggests that worst-case bounds, such as (5), may be overly conservative. To reconcile this discrepancy, it is crucial to bear in mind the intrinsic dimension of the target data distribution and explore whether and how diffusion models can harness this potentially low-dimensional structure.

The development of diffusion model theory that can effectively account for low dimensionality is, however, still in its early stages. For example, the ability of DDIM to adapt to low-dimensional structure was previously out of reach in theory, despite its widespread use. The situation for DDPM is more advanced: a few recent papers (e.g., Li and Yan (2024a); Azangulov et al. (2024); Potaptchik et al. (2024); Huang et al. (2024b)) explored its low-dimensional adaptation capability, assuming that the target data distribution is supported on some low-dimensional structure like a manifold. These studies focused primarily on convergence in Kullback–Leibler (KL) divergence, which, as we shall explain momentarily, is known to yield loose results when directly translated into convergence guarantees based on other metrics like the TV distance.

1.3 This paper

An overview of our contributions. In this paper, we develop a new suite of total-variation-based convergence guarantees for the DDIM and DDPM samplers, aimed at uncovering how they leverage low-dimensional structure to accelerate sampling. More concretely, consider a general definition of intrinsic dimension for the target distribution p_{data}, such that the intrinsic dimension is k if the logarithm of the covering number of the support of p_{data} is on the order of k (up to some log factor). With this type of intrinsic dimension in mind, we prove in Theorems 1-3 that both DDPM and DDIM take no more than the order of

$$\frac{k}{\varepsilon} \text{ iterations} \quad \text{(up to log factor)} \tag{6}$$

to generate a sample that is ε-close in TV distance to the target distribution, assuming availability of perfect score estimates. Note that we do not impose stringent assumptions like smoothness or log-concavity on p_{data}. For those applications where $k \ll d$ — a situation that is prevalent in many modern-day applications — our theory underscores the striking capability of diffusion models to automatically exploit the favorable intrinsic structure of p_{data} without explicitly modeling the low-dimensional structure or altering the algorithms. Importantly, these results provide the first theory justifying the low-dimensional adaptation ability of the DDIM-type samplers, and significantly improve over the state-of-the-art DDPM theory regarding total variation convergence; see Table 1 and Table 2 for detailed comparisons with prior DDIM and DDPM theory, respectively.[1] These convergence guarantees are also shown to be robust vis-à-vis ℓ_2 score estimation errors. Furthermore, we illuminate the specific coefficient choices of the DDIM/DDPM samplers, by linking them with reverse-time differential equations with specific discretization to exploit low dimensionality. Finally, we develop a lower bound for a single step of the discretized reverse process, which unravels the necessity and optimality of the coefficient designs proposed originally by Ho et al. (2020); Song et al. (2020).

Notation. For any positive integer n, let $[n] = \{1, \ldots, n\}$. For any two functions f and g, we employ the notation $f = O(g)$ or $f \lesssim g$ to mean that there exists some universal constant $C > 0$ such that $f \leq Cg$. The notation $f = \widetilde{O}(g)$ is defined analogously except that the logarithmic dependency is hidden. Additionally, $f \gtrsim g$ means $g \lesssim f$, and $f \asymp g$ means $f \lesssim g$ and $g \lesssim f$ hold at once. For any two distributions p and q, we denote by $\mathsf{TV}(p, q)$ (resp. $\mathsf{KL}(p \parallel q)$) the TV distance between p and q (resp. the KL divergence from q to p). We denote by p_{X_t} and p_{Y_t} the probability density function of X_t and Y_t, respectively. For any matrix A, we denote by $\|A\|$ (resp. $\|A\|_{\mathrm{F}}$) the spectral norm (resp. Frobenius norm) of A, and $\mathrm{tr}(A)$ the trace of A. For any vector-valued function $f(x)$, we let $\frac{\partial f}{\partial x}$ represent the Jacobian matrix of $f(x)$; for any real-valued

[1] Note that in a large fraction of prior DDPM theory, the bound based on the TV distance is obtained by applying Pinsker's inequality (i.e., $\mathsf{TV}(p_{X_1}, q_{Y_1}) \leq \sqrt{2\mathsf{KL}(p_{X_1} \parallel p_{Y_1})}$).

paper	smoothness of scores	score matching assumption	convergence rate (in total variation)	iteration complexity	adaptation to low dimension
Chen et al. (2023d)	L-Lipschitz	$s_t = s_t^*$	$\mathrm{poly}(Ld)/\sqrt{T}$	$\mathrm{poly}(Ld)/\varepsilon^2$	✗
Li et al. (2023a)	no requirement	$s_t \approx s_t^*$, $\frac{\partial s_t}{\partial x} \approx \frac{\partial s_t^*}{\partial x}$	$d^2/T + d^6/T^2$	$d^2/\varepsilon + d^3/\sqrt{\varepsilon}$	✗
Huang et al. (2024a)	L-Lipschitz	$s_t \approx s_t^*$	$L^2 d^2/T$	$L^2 d^2/\varepsilon$	✗
Li et al. (2024c)	no requirement	$s_t \approx s_t^*$, $\frac{\partial s_t}{\partial x} \approx \frac{\partial s_t^*}{\partial x}$	d/T when $T > d^2$	$d/\varepsilon + d^2$	✗
Li et al. (2024d)	L-Lipschitz	$s_t \approx s_t^*$, $\frac{\partial s_t}{\partial x} \approx \frac{\partial s_t^*}{\partial x}$	$Ld(L+d)/T$	$Ld(L+d)/\varepsilon$	✗
Our work (Theorem 1)	no requirement	$s_t \approx s_t^*$, $\frac{\partial s_t}{\partial x} \approx \frac{\partial s_t^*}{\partial x}$, $\nabla \mathrm{tr}(\frac{\partial s_t}{\partial x}) \approx \nabla \mathrm{tr}(\frac{\partial s_t^*}{\partial x})$	k/T	k/ε	✓

Table 1: Comparison with prior DDIM theory. The convergence rates and iteration complexities provided here assume accurate scores and ignore log factors, where the iteration complexity refers to the number of iterations needed to yield ε precision in total variation.

function $g(x)$, we let $\nabla g(x)$ represent the gradient of $g(x)$. Also, for any random object X, we denote by $\mathsf{supp}(X)$ the support of X.

2 Preliminaries

Before proceeding to our formal theory and analysis, we briefly overview some basics and the operational mechanism of diffusion models, covering both DDIM and DDPM.

Forward process and noise schedule. As previously described in (1), the forward process progressively injects Gaussian noise to transform the target distribution p_{data} into a pure noise distribution that is easy to sample from. The Gaussian nature of the injected noise allows for a more direct relation between X_0 and X_t as follows:

$$X_t = \sqrt{\overline{\alpha}_t} X_0 + \sqrt{1-\overline{\alpha}_t}\, \overline{W}_t \qquad \text{with } \overline{W}_t \sim \mathcal{N}(0, I_d), \tag{7}$$

where we introduce the following parameters for any $1 \le t \le T$:

$$\overline{\alpha}_t := \prod_{i=1}^{t} \alpha_i. \tag{8}$$

As it turns out, the choices of the coefficients $\{\alpha_t\}$ play an important role in determining the convergence properties of diffusion models. Here and throughout, we adopt the choices used in the previous work (Li et al., 2024c; Li and Yan, 2024a,b):

$$\begin{aligned} \beta_1 &:= 1 - \alpha_1 = \frac{1}{T^{c_0}}, \\ \beta_{t+1} &:= 1 - \alpha_{t+1} = \frac{c_1 \log T}{T} \min\left\{ \beta_1 \left(1 + \frac{c_1 \log T}{T}\right)^t, 1 \right\}, \quad 1 \le t < T, \end{aligned} \tag{9}$$

where $c_0, c_1 > 0$ are some large enough numerical constants. In words, this noise variance schedule (as β_t is the variance of the noise injected at step t) contains two phases: it grows exponentially at the beginning, and then stays flat after reaching the order of $\frac{\log T}{T}$, which is consistent with the state-of-the-art diffusion model theory (e.g., Benton et al. (2024); Potaptchik et al. (2024); Huang et al. (2024b); Li and Yan (2024b); Li et al. (2024c)).

paper	smoothness of scores	score matching assumption	convergence rate (in total variation)	iteration complexity	adaptation to low dimension
Chen et al. (2022b)	L-Lipschitz	$s_t \approx s_t^*$	$L\sqrt{d/T}$	$L^2 d/\varepsilon^2$	✗
Lee et al. (2022)	no requirement	$s_t \approx s_t^*$	$\sqrt{d^3/T}$	d^3/ε^2	✗
Chen et al. (2023a)	no requirement	$s_t \approx s_t^*$	$\sqrt{d^2/T}$	d^2/ε^2	✗
Benton et al. (2024)	no requirement	$s_t \approx s_t^*$	$\sqrt{d/T}$	d/ε^2	✗
Liang et al. (2024)	no requirement	$s_t \approx s_t^*$, $\nabla s_t \approx \nabla s_t^*$	$d^{3/2}/T$	$d^{3/2}/\varepsilon$	✗
Li and Yan (2024a)	no requirement	$s_t \approx s_t^*$	$k^2/\sqrt{T}$	k^4/ε^2	✓
Li and Yan (2024b)	no requirement	$s_t \approx s_t^*$	d/T	d/ε	✗
Azangulov et al. (2024)	no requirement	$s_t \approx s_t^*$	$\sqrt{k^3/T}$	k^3/ε^2	✓
Potaptchik et al. (2024)	no requirement	$s_t \approx s_t^*$	k/ε^2	$\sqrt{k/T}$	✓
Huang et al. (2024b)	no requirement	$s_t \approx s_t^*$	k/ε^2	$\sqrt{k/T}$	✓
Our work (Theorems 2-3)	no requirement	$s_t \approx s_t^*$	k/T	k/ε	✓

Table 2: Comparison with prior DDPM theory. The convergence rates and iteration complexities provided here assume accurate scores and ignore log factors, where the iteration complexity refers to the number of iterations needed to yield ε accuracy in total-variation distance.

Score-based generative models. Next, we describe the precise update rules for both DDIM-type and DDPM-type samplers.

- *DDIM-type samplers.* As mentioned previously, a DDIM-type sampler starts with $Y_T \sim \mathcal{N}(0, I_d)$ and adopts the following deterministic update rule:

$$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\big(Y_t + \eta_t^{\mathsf{ddim}} s_t(Y_t)\big), \qquad t = T, \cdots, 1. \tag{10}$$

  Here, η_t^{ddim} is a design parameter that admits multiple alternatives, and we list a couple of choices used in previous literature:

$$\eta_t^{\mathsf{ddim}} = \begin{cases} \dfrac{1-\alpha_t}{1+\sqrt{\frac{\alpha_t - \overline{\alpha}_t}{1-\overline{\alpha}_t}}} & \text{(original DDIM (Song et al., 2020); this work)} \quad (11a) \\ \dfrac{1-\alpha_t}{2} & \text{(Li et al., 2023a, 2024d,c)} \quad (11b) \\ \dfrac{-1+4\sqrt{\alpha_t}-3\alpha_t}{2\sqrt{\alpha_t}} & \text{(Song et al., 2021)} \quad (11c) \end{cases}$$

  Importantly, all of these parameter choices lead to samplers that are asymptotically consistent — meaning that the distribution of the sampling output converges to the target data distribution as T grows — under mild conditions on the target data distribution. In this paper, we concentrate on the parameter schedule (11a) proposed in the original DDIM paper (Song et al., 2020).

- *DDPM-type samplers.* A DDPM-type sampler adopts the initialization $Y_T \sim \mathcal{N}(0, I_d)$ and implements the stochastic update rule:

$$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\big(Y_t + \eta_t^{\mathsf{ddpm}} s_t(Y_t) + \sigma_t^{\mathsf{ddpm}} Z_t\big), \qquad t = T, \cdots, 1, \tag{12}$$

with independent noise $Z_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$. Here, η_t^{ddpm} and σ_t^{ddpm} are design parameters, with several choices listed below:

$$(\eta_t^{\mathsf{ddpm}}, \sigma_t^{\mathsf{ddpm}}) = \begin{cases} \left(1-\alpha_t, \sqrt{\dfrac{(1-\alpha_t)(\alpha_t - \overline{\alpha}_t)}{1-\overline{\alpha}_t}}\right) & \left(\begin{array}{c}\text{original DDPM (Ho et al., 2020);} \\ \text{Potaptchik et al. (2024);} \\ \text{Huang et al. (2024b);} \\ \text{a special case of } \underline{\text{this work}}\end{array}\right) & (13a) \\ \left(2(1-\sqrt{\alpha_t}), \sqrt{1-\alpha_t}\right) & \text{(Benton et al., 2024; Chen et al., 2023a)} & (13b) \\ \left(1-\alpha_t, \sqrt{1-\alpha_t}\right) & \text{(Li et al., 2023a; Li and Yan, 2024b)} & (13c) \end{cases}$$

All of the above choices come with convergence theory guaranteeing asymptotic consistency. In this work, we would like to accommodate a range of parameter schedules that subsumes as a special case the one (13a) proposed in the original DDPM paper (Ho et al., 2020).

ODE and SDE perspectives. To shed light on the rationale and feasibility of the DDIM-type and DDPM-type samplers, it is helpful to look at the continuous-time analogs of both forward and backward processes and resort to the toolbox of ordinary differential equations (ODEs) and stochastic differential equations (SDEs). We briefly review some basics in the sequel, and will illuminate deeper connections in Section 3.4.

- *Forward SDE.* The forward process (7) is intimately connected with the following continuous-time process with some specific choice of $\beta(t)$:

$$\mathrm{d}X_t = -\beta(t) X_t \mathrm{d}t + \sqrt{2\beta(t)}\,\mathrm{d}B_t, \tag{14}$$

where (B_t) represents a standard Brownian motion in $\mathbb{R}^d$. In fact, standard SDE theory reveals that SDE (14) admits the following characterization

$$X_t = \exp\left(-\int_0^t \beta(s)\mathrm{d}s\right) X_0 + \sqrt{1 - \exp\left(-2\int_0^t \beta(s)\mathrm{d}s\right)}\,\overline{W}_t \tag{15}$$

for some $\overline{W}_t \sim \mathcal{N}(0, I_d)$, whose marginal distribution coincides with that of Eqn. (7) if we set

$$\alpha_t = \exp\left(-2\int_{t-1}^t \beta(s)\mathrm{d}s\right), \qquad t = 1, \cdots, T. \tag{16}$$

- *Probability flow ODE or diffusion ODE.* One way to reverse the forward process is through the so-called probability flow ODE (Song et al., 2021) (also known as diffusion ODE):

$$\mathrm{d}Y_t = \left(Y_t + s^*_{T-t}(Y_t)\right)\beta(T-t)\mathrm{d}t, \qquad t \in [0, T], \tag{17}$$

which enjoys matching marginal distribution $Y_{T-t} \overset{\mathrm{d}}{=} X_t$ for all $0 \le t \le T$ as long as we generate $Y_0 \sim p_{X_T}$. To approximately simulate this reverse ODE in practice and obtain a tractable sampler, a common strategy is to perform time discretization of ODE (17). Note that different discretization schemes can result in different design coefficients η_t^{ddim} as in the DDIM-type update rule (10).

- *Reverse-time SDE.* An alternative way to reverse the forward process is via a properly chosen SDE. In view of the classical results in the SDE literature (Anderson, 1982; Haussmann and Pardoux, 1986), the following SDE,

$$\mathrm{d}Y_t = \left(Y_t + 2s^*_{T-t}(Y_t)\right)\beta(T-t)\mathrm{d}t + \sqrt{2\beta(T-t)}\,\mathrm{d}W_t, \qquad t \in [0, T] \tag{18}$$

with (W_t) a standard Brownian motion in $\mathbb{R}^d$, reverses the forward process (14) in the sense that $Y_{T-t} \overset{\mathrm{d}}{=} X_t$ for all $0 \le t \le T$ as long as $Y_0 \sim p_{X_T}$. Akin to the DDIM counterpart, the DDPM-type samplers can often be viewed as time discretization of SDE (18), and different discretization schemes correspond to different coefficient choices of $(\eta_t^{\mathsf{ddpm}}, \sigma_t^{\mathsf{ddpm}})$ as in (12).

- *Generalized reverse-time ODE/SDE.*

$$\mathrm{d}Y_t = \big(Y_t + \big(1+\xi(T-t)\big)s^*_{T-t}(Y_t)\big)\,\beta(T-t)\mathrm{d}t + \sqrt{2\xi(T-t)\beta(T-t)}\,\mathrm{d}W_t, \qquad t\in[0,T] \tag{19}$$

for some general function $\xi(t)\geq 0$ for all $0\leq t\leq T$, where (W_t) again represents a standard Brownian motion in $\mathbb{R}^d$. We shall formally demonstrate the desired distributional property of this family of differential equations in Appendix E.1. When $\xi(t)=0$ (resp. $\xi(t)=1$) for all $t\in[0,T]$, (19) reduces to ODE (17) (resp. SDE (18)). For a general $\xi(t)$, suitable time discretizationschemes of (19) can lead to new samplers other than the original DDIM and DDPM.

3 Main results

In this section, we present our main results and discuss their implications. The key assumptions are introduced in Section 3.1, followed by our convergence theory in Sections 3.2-3.3.

3.1 Key assumptions

To begin with, let us single out two assumptions concerning the target data distribution p_{data}. We denote by $\mathcal{X}_{\mathsf{data}}\in\mathbb{R}^d$ the support of p_{data}, i.e., the closure of the intersection of all the sets $\mathcal{X}'\in\mathbb{R}^d$ such that $\mathbb{P}_{X_0\sim p_{\mathsf{data}}}(X_0\in\mathcal{X}')=1$. In order to rigorously define the "intrinsic dimension" of p_{data}, we find it convenient to introduce the following definition of covering number (Wainwright, 2019, Chapter 5), which provides a generic way to measure the complexity of a set $\mathcal{X}$.

Definition 1 (Covering number) *For any set $\mathcal{X}\subseteq\mathbb{R}^d$, the (Euclidean) covering number at scale $\epsilon_0>0$, denoted by $N_{\epsilon_0}(\mathcal{X})$, is defined as the smallest integer n such that there exist points $x_1,\ldots,x_n$ obeying*

$$\mathcal{X}_{\mathsf{data}}\subseteq\bigcup_{i=1}^{n}\mathcal{B}(x_i,\epsilon_0),$$

where $\mathcal{B}(x_i,\epsilon_0):=\{x\in\mathbb{R}^d\mid\|x-x_i\|_2\leq\epsilon_0\}$ and $\|\cdot\|_2$ denotes the ℓ_2 norm.

The covering number in turn enables a flexible characterization of the complexity of the data distribution.

Assumption 1 (Intrinsic dimension) *Consider $\epsilon_0=T^{-c_{\epsilon_0}}$ for some sufficiently large universal constant $c_{\epsilon_0}>0$. The covering number of the support $\mathcal{X}_{\mathsf{data}}$ of p_{data} is assumed to satisfy*

$$\log N_{\epsilon_0}(\mathcal{X}_{\mathsf{data}})\leq C_{\mathsf{cover}}k\log T$$

for some constant $C_{\mathsf{cover}}>0$. Here and throughout, we shall refer to k as the intrinsic dimension of p_{data}.

The intrinsic dimension defined above is fairly generic, facilitating studies of a number of important low-dimensional structures. Partial examples that satisfy Assumption 1 include k-dimensional linear subspace in $\mathbb{R}^d$ and k-dimensional non-linear manifolds (provided that $\mathcal{X}_{\mathsf{data}}$ is polynomially bounded as in Assumption 2 below), as well as structures with doubling dimension k (Dasgupta and Freund, 2008). The interested reader is referred to Huang et al. (2024b, Section 4.1) for a more detailed discussion.

The second assumption we would like to impose on p_{data} is the boundedness of its support as follows.

Assumption 2 (Bounded support) *Suppose that there exists a universal constant $c_R>0$ such that*

$$\sup_{x\in\mathcal{X}_{\mathsf{data}}}\|x\|_2\leq R \quad \textit{where} \quad R:=T^{c_R}.$$

Note that the size of the support $\mathcal{X}_{\mathsf{data}}$ is allowed to scale polynomially (with arbitrarily large degree) in the number of iterations of the sampler, which accommodates a very wide range of practical applications like image generation.

Next, we turn to the quality of score estimates and impose the following assumption regarding their ℓ_2 accuracy. It is noteworthy that the score error metric $\varepsilon_{\mathsf{score}}$ defined below captures the mean squared estimation error when averaged over all time steps, rather than representing the error for a single time step.

Assumption 3 (ℓ_2 score estimation error) *Suppose that the estimated score functions $\{s_t(\cdot)\}_{t=1}^T$ obey*

$$\frac{1}{T}\sum_{t=1}^{T}\varepsilon_{\mathsf{score},t}^2 \le \varepsilon_{\mathsf{score}}^2 \qquad \text{with } \varepsilon_{\mathsf{score},t}^2 := \mathbb{E}\big[\,\|s_t(X_t)-s_t^*(X_t)\|_2^2\,\big]. \tag{20}$$

The ℓ_2 score estimation error is commonly assumed in the state-of-the-art results on diffusion models (e.g., Chen et al. (2022a); Benton et al. (2024)). Additionally, this form of estimation error also aligns with practical training procedures such as score matching (e.g., Hyvärinen (2005); Vincent (2011)).

Finally, while our convergence theory for both DDIM and DDPM relies upon Assumptions 1-3, these assumptions alone are insufficient to guarantee convergence of the DDIM-type samplers; see Li et al. (2024c, Section 3.2) for a counterexample. Consequently, we introduce below an additional set of assumptions in order to establish convergence theory for DDIM-type samplers.

Assumption 4 (Additional score estimation assumption for DDIM) *Consider the estimated score functions $\{s_t(\cdot)\}_{t=1}^T$. Assume that $s_t(\cdot)$ is twice continuously differentiable for each $1\le t\le T$, and suppose*

$$\frac{1}{T}\sum_{t=1}^{T}\varepsilon_{\mathsf{Jacobi},1,t}^2 \le \varepsilon_{\mathsf{Jacobi},1}^2 \qquad \text{with } \varepsilon_{\mathsf{Jacobi},1,t}^2 := \mathbb{E}\left[\left\|\frac{\partial s_t(X_t)}{\partial x}-\frac{\partial s_t^*(X_t)}{\partial x}\right\|_{\mathrm{F}}^2\right], \tag{21a}$$

$$\frac{1}{T}\sum_{t=1}^{T}\varepsilon_{\mathsf{Jacobi},2,t}^2 \le \varepsilon_{\mathsf{Jacobi},2}^2 \qquad \text{with } \varepsilon_{\mathsf{Jacobi},2,t}^2 := \mathbb{E}\left[\mathrm{tr}\left(\frac{\partial s_t(X_t)}{\partial x}-\frac{\partial s_t^*(X_t)}{\partial x}\right)^2\right], \tag{21b}$$

$$\frac{1}{T}\sum_{t=1}^{T}\varepsilon_{\mathsf{Hess},t}^2 \le \varepsilon_{\mathsf{Hess}}^2 \qquad \text{with } \varepsilon_{\mathsf{Hess},t}^2 := \mathbb{E}\left[\left\|\nabla\mathrm{tr}\left(\frac{\partial s_t(X_t)}{\partial x}-\frac{\partial s_t^*(X_t)}{\partial x}\right)\right\|_2^2\right]. \tag{21c}$$

In short, Assumption 4 is concerned with higher-order estimation errors of the score functions under different metrics, namely, the time-averaged errors w.r.t. the associated Jacobian matrix and Hessian tensor. Intuitively, given that DDIM is deterministic without bringing in random noise to smooth the trajectory, additional assumptions like higher-order score estimation accuracy are needed in order to mitigate the propagation of estimation errors in each backward step. We shall see how these error metrics influence the final sampling fidelity in the next subsection.

3.2 Convergence theory for DDIM

We are now positioned to present below our total-variation-based convergence theory for the DDIM sampler in the presence of low-dimensional structure. The proof of this theorem is postponed to Appendix B.

Theorem 1 *Under Assumptions 1-4, the DDIM sampler (3) with the coefficients $\eta_t^{\mathsf{ddim}} = \frac{1-\alpha_t}{1+\sqrt{\frac{\alpha_t-\bar{\alpha}_t}{1-\bar{\alpha}_t}}}$ yields*

$$\mathsf{TV}(p_{X_1},p_{Y_1}) \lesssim \frac{k\log^3 T}{T} + (\varepsilon_{\mathsf{score}}+\varepsilon_{\mathsf{Jacobi},1}+\varepsilon_{\mathsf{Jacobi},2}+\varepsilon_{\mathsf{Hess}})\sqrt{\log T}. \tag{22}$$

To the best of our knowledge, this provides the first theory that unveils how the DDIM sampler adapts to unknown low-dimensional structure of p_{data}; see Table 1 for a summary of prior results. Noteworthily, this convergence theory accommodates a very broad family of target data distributions p_{data}, without requiring stringent assumptions like smoothness or log-concavity. Several remarks are in order.

- *Iteration complexity.* When accurate scores (i.e., $s_t = s_t^*$ for all t) are available, the number of steps needed for the DDIM sampler to achieve $\mathsf{TV}(p_{X_1},p_{Y_1})\le\varepsilon$ scales as

$$\widetilde{O}\left(\frac{k}{\varepsilon}\right). \tag{23}$$

  As a consequence, if the intrinsic dimension $k \ll d$, then the DDIM sampler automatically accelerates, without any prior knowledge about the underlying low-dimensional structure.

- *No burn-in cost.* We also compare Theorem 1 with the state-of-the-art DDIM theory (Li et al., 2024c) for the case with $k = d$ and accurate scores. Recall that Li et al. (2024c) established an $\widetilde{O}(d/T)$ convergence rate, which is consistent with Theorem 1. Nevertheless, the theory therein requires a burn-in cost $T \gtrsim d^2 \log^5 T$, a condition that contrasts sharply with our theorem as we do not impose such a burn-in requirement.

- *Coefficient choices.* Interestingly, the remarkable low-dimensional adaptation capability is achieved with the coefficient $\eta_t^{\mathsf{ddim}} = \frac{1-\alpha_t}{1+\sqrt{\frac{\alpha_t - \overline{\alpha}_t}{1-\overline{\alpha}_t}}}$, which matches exactly the coefficient proposed for the original DDIM sampler (Song et al., 2020). As we shall elaborate on momentarily, not all the coefficient choices in (11) are capable of adapting to low-dimensional structure.

- *Second-order assumptions on score estimation.* Unlike previous convergence analysis for DDIM, Theorem 1 makes an assumption about the second-order approximation of $s_t(\cdot)$ to $s_t^*(\cdot)$, i.e., the additional error term $\varepsilon_{\mathsf{Hess}}$ in Assumption 4. This arises because, in prior studies, the score error terms in the convergence rate of DDIM were dependent on the ambient dimension d. For instance, in Huang et al. (2024a); Li et al. (2024c), the estimation error terms in their respective convergence rates are given by $d^{\frac{3}{4}} L^{\frac{1}{2}} \varepsilon_{\mathsf{score}}$ and $\sqrt{d}\varepsilon_{\mathsf{score}} + d\varepsilon_{\mathsf{Jacobi}}$. In comparison, in our Theorem 1, the score estimation error term in the sampling error is nearly dimension-free (except for logarithmic dependency), meaning that this error does not amplify when the intrinsic and ambient dimensions increase.

3.3 Convergence theory for DDPM

Turning attention to the DDPM-type samplers, we present below our total-variation-based convergence guarantees for the original DDPM sampler proposed by Ho et al. (2020). The proof can be found in Appendix C.

Theorem 2 *Under Assumptions 1-3, the DDPM sampler (4) with the coefficients* $\eta_t^{\mathsf{ddpm}} = 1 - \alpha_t$ *and* $\sigma_t^{\mathsf{ddpm}} = \sqrt{\frac{(\alpha_t - \overline{\alpha}_t)(1-\alpha_t)}{1-\overline{\alpha}_t}}$ *achieves*

$$\mathsf{TV}(p_{X_1}, p_{Y_1}) \lesssim \frac{k \log^3 T}{T} + \varepsilon_{\mathsf{score}} \sqrt{\log T}. \tag{24}$$

Akin to our DDIM theory, the DDPM sampler — using coefficients proposed in the original DDPM work Ho et al. (2020) — achieves an iteration complexity no greater than

$$\widetilde{O}\left(\frac{k}{\varepsilon}\right) \tag{25}$$

when exact score estimates are available, without requiring any sort of smoothness or log-concavity assumptions. Our result improves upon the state-of-the-art general theory for DDPM (i.e., $\widetilde{O}(d/\varepsilon)$ as established by Li and Yan (2024b)) by a factor of d/k, uncovering a substantial speed-up when $k \ll d$. It is worth noting that low-dimensional adaptation of the DDPM sampler was first rigorized by Li and Yan (2024a), followed by a couple of recent papers to sharpen the KL-based convergence guarantees (Azangulov et al., 2024; Potaptchik et al., 2024; Huang et al., 2024b). Nevertheless, directly combining Pinsker's inequality with these KL-based bounds falls short of delivering tight TV-based results. See Table 2 for more detailed comparisons.

Additionally, when score estimation is imperfect, our TV-based convergence guarantees degrade gracefully, with the bounds scaling linearly in $\varepsilon_{\mathsf{score}}$ (a metric that measures the ℓ_2 estimation error). In stark contrast to our DDIM theory in Theorem 1, the convergence of DDPM can be established under fewer assumptions; for instance, there is no need of imposing assumptions on the Jacobian or Hessian of score estimates as in Assumption 4. This favorable feature of DDPM arises since its stochastic update rule introduces additional Gaussian noise in each step, which helps smooth the trajectory and eliminates the need to cope with many boundary cases.

As it turns out, the coefficients (13a) are not the only choice of DDPM-type samplers that enable the desirable adaptation. Our convergence theory can be extended to accommodate a broader set of coefficients, as summarized in the following theorem. The proof is deferred to Section C.

Theorem 3 *Suppose that the coefficients η_t^{ddpm} and σ_t^{ddpm} satisfy*

$$(1-\overline{\alpha}_t)\left(1-\frac{\eta_t^{\mathsf{ddpm}}}{1-\overline{\alpha}_t}\right)^2 = \alpha_t - \overline{\alpha}_t - \left(\sigma_t^{\mathsf{ddpm}}\right)^2, \qquad t=1,\dots,T. \tag{26}$$

Also, assume that there exists some universal constant $C_1 \geq 1/2$ such that

$$\eta_t^{\mathsf{ddpm}} \leq \min\left\{C_1(1-\alpha_t), \frac{1}{2}(1-\overline{\alpha}_t)\right\}, \qquad t=1,\cdots,T. \tag{27}$$

- *Consider the case with exact score estimation, i.e., $s_t = s_t^*$ for all $t=1,\dots,T$. Then under Assumptions 1-2, the DDPM sampler (4) yields*

$$\mathsf{TV}(p_{X_1}, p_{Y_1}) \lesssim \frac{k\log^3 T}{T}.$$

- *Consider the case with imperfect score estimation. Also, assume that*

$$(\eta_t^{\mathsf{ddpm}})^2 \leq C_2(1-\alpha_t)(\sigma_t^{\mathsf{ddpm}})^2, \qquad t=1,\cdots,T \tag{28}$$

  for some universal constant $C_2 > 0$. Then under Assumptions 1-3, the DDPM sampler (4) yields

$$\mathsf{TV}(p_{X_1}, p_{Y_1}) \lesssim \frac{k\log^3 T}{T} + \varepsilon_{\mathsf{score}}\sqrt{\log T}.$$

Let us take a moment to discuss the range of coefficients satisfying relation (26). Interestingly, this relation between η_t^{ddpm} and σ_t^{ddpm} aligns perfectly with the set of coefficients discussed in Song et al. (2020, Section 4.1). More specifically, Song et al. (2020, Eq. (12)) singled out the update rule below:

$$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(Y_t - \sqrt{1-\overline{\alpha}_t}\epsilon_t^{\mathsf{noise}}(Y_t) + \sqrt{\alpha_t - \overline{\alpha}_t - \alpha_t\varsigma_t^2}\epsilon_t^{\mathsf{noise}}(Y_t) + \sqrt{\alpha_t}\varsigma_t Z_t\right) \tag{29}$$

for some coefficient ς_t, where $\epsilon_t^{\mathsf{noise}}(Y_t) = -\sqrt{1-\overline{\alpha}_t}s_t(Y_t)$ serves as an estimate of the noise injected in the forward process. One can demonstrate its equivalence with (26). The interested reader is referred to Appendix D for more details.

Additionally, it was conjectured in Li and Yan (2024a) that the coefficients studied therein might not be the only optimal choice when it comes to total-variation convergence. In light of this, Theorem 3 addresses this conjecture by showing that the coefficients analyzed therein are a special case of a wider range of feasible coefficients.

3.4 Interpretation from the lens of differential equations

In order to help elucidate why DDIM and DDPM are adaptive to low dimensionality, we take a moment to derive their exact correspondence to reversed differential equations. This viewpoint unearths the underlying forces that steer their trajectories toward the low-dimensional structure of interest, despite the effects of time discretization. For convenience of presentation, we overload the notation by setting

$$\overline{\alpha}_t = \exp\left(-2\int_0^t \beta(s)\mathrm{d}s\right), \qquad \text{for all } t\in[0,T], \tag{30}$$

where $\beta(t)$ denotes the coefficient schedule in the forward SDE (14); as alluded to previously, this function $\overline{\alpha}_t$ coincides with $\{\overline{\alpha}_t\}_{t=1}^T$ defined in (8) for the discrete-time process. In addition, we recall from the Tweedie formula (Efron, 2011) that

$$\mu_t^*(x) \coloneqq \mathbb{E}[X_0 \mid X_t = x] = \frac{1}{\sqrt{\overline{\alpha}_t}}\big(x + (1-\overline{\alpha}_t)s_t^*(x)\big), \tag{31a}$$

and also introduce the noisy counterpart:

$$\mu_t(x) \coloneqq \frac{1}{\sqrt{\overline{\alpha}_t}}\big(x + (1-\overline{\alpha}_t)s_t(x)\big). \tag{31b}$$

In the sequel, we isolate several discretized differential equations that correpond exactly with the DDIM and DDPM samplers considered in the present paper.

- *DDIM sampler.* The probability flow ODE (17) can be reparametrized by μ_t^* as follows using Tweedie's formula (31a):

$$\mathrm{d}Y_t = \left(-\frac{\overline{\alpha}_{T-t}}{1-\overline{\alpha}_{T-t}}Y_t + \frac{\sqrt{\overline{\alpha}_{T-t}}}{1-\overline{\alpha}_{T-t}}\mu_{T-t}^*(Y_t)\right)\beta(T-t)\mathrm{d}t,$$

  where the drift term exhibits a semi-linear structure. To approximately solve this ODE, one can apply the exponential integrator scheme on the estimated semi-linear structure and select time discretization points as $t_n = n$ for all $n = 0, 1, \cdots, T$, leading to the discretized dynamics below:

$$\mathrm{d}\widetilde{Y}_t = \left(-\frac{\overline{\alpha}_{T-t}}{1-\overline{\alpha}_{T-t}}\widetilde{Y}_t + \frac{\sqrt{\overline{\alpha}_{T-t}}}{1-\overline{\alpha}_{T-t}}\mu_{T-t_n}(\widetilde{Y}_{t_n})\right)\beta(T-t)\mathrm{d}t, \quad t \in [t_n, t_{n+1}). \tag{32}$$

  The DDIM sampler is intimately connected with this discretized dynamic, as asserted by the following proposition, whose proof can be found in Appendix E.3.

  Proposition 1 *The discretized process (32) is solved exactly by the DDIM update rule (3) with coefficient (11a) in the sense that $\widetilde{Y}_n = Y_{T-n}$ for all $n = 0, 1, \ldots, T$, provided that $\widetilde{Y}_0 = Y_T$.*

- *(Generalized) DDPM sampler.* Similarly, the DDPM-type sampler — with the coefficients chosen as in Theorem 3 — can be exactly mapped to certain discretized differential equations. More precisely, consider the generalized semi-linear SDE/ODE (19), which can be reparametrized via μ_t^* through Tweedie's formula (31a):

$$\mathrm{d}Y_t = \left(-\frac{\xi(T-t)+\overline{\alpha}_{T-t}}{1-\overline{\alpha}_{T-t}}Y_t + \frac{(1+\xi(T-t))\sqrt{\overline{\alpha}_{T-t}}}{1-\overline{\alpha}_{T-t}}\mu_{T-t}^*(Y_t)\right)\beta(T-t)\mathrm{d}t + \sqrt{2\xi(Y-t)\beta(T-t)}\,\mathrm{d}W_t.$$

  Adopting similar discretization scheme as in (32), we arrive at the following discretized process:

$$\begin{aligned}\mathrm{d}\widetilde{Y}_t = \left(-\frac{\xi(T-t_n)+\overline{\alpha}_{T-t}}{1-\overline{\alpha}_{T-t}}\widetilde{Y}_t + \frac{(1+\xi(T-t_n))\sqrt{\overline{\alpha}_{T-t}}}{1-\overline{\alpha}_{T-t}}\mu_{T-t_n}(\widetilde{Y}_{t_n})\right)\beta(T-t)\mathrm{d}t \\ + \sqrt{2\xi(T-t_n)\beta(T-t)}\,\mathrm{d}W_t, \quad t \in [t_n, t_{n+1}),\end{aligned} \tag{33}$$

  where we recall that $t_n = n$. Interestingly, the DDPM-type samplers considered in Theorem 3 correspond exactly to (33) with suitably chosen $\xi(t)$, as stated below. The proof of Proposition 2 can be found in Appendix E.2.

  Proposition 2 *The discretized process (33) can be solved exactly by the (generalized) DDPM update rule (4) — with coefficients satisfying (26) — in the sense that $\widetilde{Y}_n = Y_{T-n}$ for all $n = 0, 1, \ldots, T$, provided that $\widetilde{Y}_0 = Y_T$ and that the standard Gaussian vectors $\{Z_t\}$ are chosen properly based on (W_t).*

Remark 1 *In the special case with coefficients (13a), the precise connection between such a discretized differential equation and the original DDPM sampler (Ho et al., 2020) has been discussed and utilized in the recent work Azangulov et al. (2024); Potaptchik et al. (2024); Huang et al. (2024b).*

With the equivalent description (32) (resp. (33)) of the DDIM (resp. DDPM) sampler, one can already gain insight into how these samplers adapt to unknown low-dimensional structure. Suppose that we have access to accurate scores, so that $\mu_t^* = \mu_t$. A closer inspection of (32) and (33) reveals that: the nonlinear components of the drift terms of both processes are proportional to $\mu_{T-t_n}^*$, which is defined as the conditional expectation of X_0 (cf. (31a)). In other words, the most critical drift components take the form of conditional expectation of X_0, which inherently capture the low-dimensional structure of p_{data} and steer the sampling dynamics towards this inherent structure.

3.5 Other alternatives of coefficient design?

Thus far, our main theorems (i.e., Theorems 1-3) focus attention on specific coefficient choices as in the original DDIM and DDPM samplers. One might naturally wonder whether other coefficient choices could also facilitate low-dimensional adaptation capabilities. As it turns out, these particular coefficients — or those exceedingly close to them — are almost necessary to achieve adaptivity, as explained in this subsection.

For simplicity, consider the case with accurate score estimates (i.e., $s_t^* = s_t$ for all t), and let us look at the following mapping:

$$\Phi_t^*(x, z) := \frac{1}{\sqrt{\alpha_t}}\big(x + \eta_t s_t^*(x) + \sigma_t z\big). \tag{34}$$

Clearly, both the DDIM update rule (10) and DDPM update rule (12) in the t-th iteration can be described as $Y_{t-1} = \Phi_t^*(Y_t, Z_t)$ for some choices of η_t and σ_t (i.e., $\sigma_t = 0$ for DDIM and $\sigma_t \neq 0$ for DDPM), where Z_t is an independent standard Gaussian vector. To evaluate how well the efficacy of DDIM-type and DDPM-type samplers, we propose to perform a sort of *one-step analysis* as follows:

1) Start the sampler from X_t of the forward process (1);
2) Compute one iteration $Y_{t-1} = \Phi_t^*(Y_t, Z_t)$ with an independent Gaussian vector $Z_t \sim \mathcal{N}(0, I_d)$;
3) Evaluate the TV distance between Y_t and X_t and see whether it is well-controlled.

An ideal sampler that can effectively adapt to unknown low dimensionality would not incur a TV distance blowing up with the ambient dimension d.

As it turns out, in order for the TV distance between X_t and Y_t to be well-controlled, the coefficients (η_t, σ_t) must be carefully chosen, as revealed by the following lower bound. The proof of this lower bound is provided in Appendix F.

Theorem 4 *Consider any $k \le d/2$, and take the target distribution p_{data} to be $\mathcal{N}\left(0, \begin{bmatrix} I_k & \\ & 0 \end{bmatrix}\right)$. Then for arbitrary choices of (η_t, σ_t), we have*

$$\mathsf{TV}\big(\Phi_t^*(X_t, Z_t), X_{t-1}\big) \ge \frac{1}{100} \min\left\{ \sqrt{\frac{d}{2}} \left| \frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}\left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right)^2 + \frac{\sigma_t^2}{\alpha_t - \overline{\alpha}_t} - 1 \right|, 1 \right\}. \tag{35}$$

In words, Theorem 4 asserts that even when initialized from a point from the true forward process, performing one iteration of DDIM/DDPM updates might already incur a TV distance that scales polynomially in the ambient dimension, unless the coefficients are chosen to obey

$$\frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}\left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right)^2 + \frac{\sigma_t^2}{\alpha_t - \overline{\alpha}_t} - 1 \approx 0. \tag{36}$$

- Consider the DDIM-type sampler (10), which has $\sigma_t = 0$. The requirement (36) then simplifies to

$$\eta_t \approx 1 - \overline{\alpha}_t - \sqrt{(\alpha_t - \overline{\alpha}_t)(1 - \overline{\alpha}_t)} = \frac{1 - \alpha_t}{1 + \sqrt{\frac{\alpha_t - \overline{\alpha}_t}{1 - \overline{\alpha}_t}}},$$

  the right-hand side of which is precisely the choice (11a) of the original DDIM sampler.

- The DDPM-type sampler (12) then corresponds to the case with $\sigma_t > 0$. Clearly, all coefficient choices studied in Theorem 3 satisfy this requirement, subsuming the choice (13a) of the original DDPM sampler as a special case.

Somewhat surprisingly, while the original DDIM and DDPM update rules (Song et al., 2020; Ho et al., 2020) were derived heuristically (namely, by maximizing some variational lower bounds on the log-likelihoods) without any explicit consideration of the low-dimensional structure, the coefficients of the resulting algorithms prove to be nearly essential for adaptation to low dimensionality.

4 Related work

General convergence analysis of diffusion models. A recent strand of work has been devoted to analyzing the convergence behavior of diffusion models (Chen et al., 2022b; Lee et al., 2022; Liu et al., 2022; Lee et al., 2023; Chen et al., 2023a; Benton et al., 2024; Chen et al., 2023d; Li et al., 2023a; Pedrotti et al., 2023; Cheng et al., 2023; Huang et al., 2024a; Liang et al., 2024; Li and Yan, 2024b; Tang and Zhao, 2024a; Li et al., 2024d; Ren et al., 2024; Gao and Zhu, 2024; Gentiloni-Silveri and Ocello, 2025); see Tang and Zhao (2024b) for a tutorial. Take the DDPM for instance, the work Chen et al. (2022b) established convergence analysis (based on Girsanov's theorem) without assuming log-concavity; the smoothness assumption was further relaxed by Lee et al. (2023); Chen et al. (2023a). Regarding the use of DDPM for a general class of non-smooth and non-log-concave distributions, Benton et al. (2024) established the best-known KL-based convergence guarantees, whereas the state-of-the-art TV-based convergence was derived by Li and Yan (2024b). Turning attention to the DDIM, Chen et al. (2023d) derived the first polynomial-time analysis, while Chen et al. (2023c) provided improved analysis for a variation of the probability flow ODE (by adding an additional stochastic step). Li et al. (2023a) improved the TV-based iteration complexity of the DDIM to $\widetilde{O}(d^2/\varepsilon)$, which was subsequently improved by Li et al. (2024c) to $\widetilde{O}(d/\varepsilon + d^2)$. Additionally, higher-order samplers tailored to solving the reverse-time SDE or probability flow ODE (e.g., Lu et al. (2022a,b)) have been proven to achieve faster convergence (Li et al., 2024a; Wu et al., 2024b; Li and Cai, 2024; Huang et al., 2024a). Randomized midpoint methods have also been leveraged to provably speed up convergence (Shen and Lee, 2019; Gupta et al., 2024; Li and Jiao, 2024). The convergence behavior of conditional diffusion models (or diffusion guidance) is another important topic that has been studied by several recent work (e.g., Wu et al. (2024a); Chidambaram et al. (2024); Chen et al. (2024a); Fu et al. (2024); Tang and Xu (2024)).

Score matching. An important stage of score-based generative modeling is score matching or score learning (Hyvärinen, 2005, 2007; Vincent, 2011; Song and Ermon, 2019; Ho et al., 2020), which aims to learn the score functions (typically using deep neural networks or transformers). From the statistical perspectives, Koehler et al. (2023) characterized the asymptotic statistical efficiency of score matching, while Oko et al. (2023); Wibisono et al. (2024); Zhang et al. (2024); Han et al. (2024); Dou et al. (2024) derived the statistical error rates and sample complexity for score matching. Another recent work Feng et al. (2024) leveraged some idea from score matching to tackle convex M-estimation. As the score matching phase is not the primary focus of our work, we do not delve into further details here.

Diffusion models in the presence of low-dimensional structure. Given the ubiquity of low-dimensional structure in practice (Pope et al., 2021), a recent line of research sought to unveil the role of low dimensionality in enabling more efficient data generation (Li and Yan, 2024a; Azangulov et al., 2024; Potaptchik

et al., 2024; Huang et al., 2024b). More concretely, Li and Yan (2024a) established the first iteration complexity upper bound for the DDPM that is adaptive to unknown low-dimensional structures, without the need of modifying the algorithm; the iteration complexity therein is proportional to k^4, with k the intrinsic dimension. This k-dependency was subsequently improved by Azangulov et al. (2024) to k^3 and then tightened by Potaptchik et al. (2024); Huang et al. (2024b) to linear scaling. However, all of these past results focused on KL-based convergence, which are loose when translated to TV-based convergence theory using Pinsker's inequality. What is more, no prior theory studied how ODE-based samplers adapt to unknown low-dimensional structure.

Apart from the above-mentioned convergence analysis for the sampling stage, the interplay between diffusion models and low-dimensional structure has been investigated from other perspectives as well (Chen et al., 2023b; Wang et al., 2024; Tang and Yang, 2024; Stanczuk et al., 2024; Mei and Wu, 2023; Li et al., 2023b; Azangulov et al., 2024; Li et al., 2024f,e; Cui et al., 2025). For instance, Chen et al. (2023b) considered the case where the target distribution lies on a linear subspace and developed sample complexity bounds for score matching that are independent of the ambient dimension. Wang et al. (2024) assumed the target distribution to be a mixture of low-rank Gaussians and explored the equivalence between score matching in this setting and subspace clustering. Tang and Yang (2024) studied the case when the data are supported on low-dimensional manifolds, and provided explicit convergence rates highlighting the importance of score estimation methods in such settings. Stanczuk et al. (2024) showed that diffusion models encode the data manifold by approximating its normal bundle. Moreover, Li et al. (2023b) established theoretical estimates of the generalization gap that evolves with the training dynamics of score-based diffusion models, suggesting a polynomially small generalization error that evades the curse of dimensionality.

5 Discussion

We have made progress towards understanding how diffusion models harness (unknown) low-dimensional structure to accelerate data generation. For the DDIM sampler (or the ODE-based sampler), we have provided the first analysis demonstrating its ability to adapt to low-dimensional structure. Along the way, we have managed to eliminate the need of a large burn-in requirement imposed in the state-of-the-art work Li et al. (2024c) for the general full-dimensional case with $k = d$. When it comes to the DDPM sampler (or the SDE-based sampler), we have improved the TV-based iteration complexity from $\widetilde{O}(k/\varepsilon^2)$ (Potaptchik et al., 2024; Huang et al., 2024b) to $\widetilde{O}(k/\varepsilon)$. It is worth noting that the coefficients analyzed in the current work align perfectly with the choices proposed originally in Ho et al. (2020); Song et al. (2020). Through a lower bound analysis, we have offered insights into the critical role of such coefficient designs in facilitating low-dimensional adaptation.

Before concluding this paper, we briefly point out a couple of directions worthy of future investigation. First, careful readers would notice that the coefficients of the DDIM analyzed in Theorem 1 is a special case of the ones satisfying (26) as isolated in Theorem 3. However, our current analysis for Theorem 3 does not yet accommodate the case when $\eta_t \gtrsim (1-\alpha_t)\sigma_t^2$, thus leaving a gap in the connection between DDIM and DDPM samplers. Bridging this gap would offer a deeper understanding of the connection between DDPM and DDIM and, potentially, offer a unified theoretical framework to study both types of samplers. Moreover, Theorem 4 currently provides only a lower bound for a single step of the discretized reverse process, which does not encompass all SDE/ODE-based samplers. It would be helpful to develop multi-step lower bounds that apply to a wider family of diffusion-based samplers. Finally, while the present paper focuses on the sampling stage, it does not unpack the score learning phase; in particular, it remains unclear how low-dimensional structure affects the efficiency of score learning. Establishing an end-to-end theory that takes into account the adaptivity of both score learning and sampling would be an avenue for future exploration.

Acknowledgments

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 grants 2221009 and 2218773. We thank Yuting Wei and Yuchen Wu for their helpful discussions.

References

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

Azangulov, I., Deligiannidis, G., and Rousseau, J. (2024). Convergence of diffusion models under the manifold hypothesis in high-dimensions. *arXiv preprint arXiv:2409.18804*.

Benton, J., Bortoli, V. D., Doucet, A., and Deligiannidis, G. (2024). Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*.

Benton, J., Deligiannidis, G., and Doucet, A. (2023). Error bounds for flow matching methods. *arXiv preprint arXiv:2305.16860*.

Chen, H., Lee, H., and Lu, J. (2022a). Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. *arXiv preprint arXiv:2211.01916*.

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

Chen, M., Huang, K., Zhao, T., and Wang, M. (2023b). Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pages 4672–4712. PMLR.

Chen, M., Mei, S., Fan, J., and Wang, M. (2024a). An overview of diffusion models: Applications, guided generation, statistical rates and optimization. *arXiv preprint arXiv:2404.07771*.

Chen, S., Chewi, S., Lee, H., Li, Y., Lu, J., and Salim, A. (2023c). The probability flow ODE is provably fast. *arXiv preprint arXiv:2305.11798*.

Chen, S., Chewi, S., Lee, H., Li, Y., Lu, J., and Salim, A. (2024b). The probability flow ode is provably fast. *Advances in Neural Information Processing Systems*, 36.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. R. (2022b). Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*.

Chen, S., Daras, G., and Dimakis, A. G. (2023d). Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for DDIM-type samplers. *arXiv preprint arXiv:2303.03384*.

Cheng, X., Lu, J., Tan, Y., and Xie, Y. (2023). Convergence of flow-based generative models via proximal gradient descent in wasserstein space. *arXiv preprint arXiv:2310.17582*.

Chidambaram, M., Gatmiry, K., Chen, S., Lee, H., and Lu, J. (2024). What does guidance do? a fine-grained analysis in a simple setting. *arXiv preprint arXiv:2409.13074*.

Croitoru, F.-A., Hondru, V., Ionescu, R. T., and Shah, M. (2023). Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869.

Cui, H., Pehlevan, C., and Lu, Y. M. (2025). A precise asymptotic analysis of learning diffusion models: theory and insights. *arXiv preprint arXiv:2501.03937*.

Dasgupta, S. and Freund, Y. (2008). Random projection trees and low dimensional manifolds. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 537–546.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Devroye, L., Mehrabian, A., and Reddad, T. (2018). The total variation distance between high-dimensional gaussians with the same mean. *arXiv preprint arXiv:1810.08693*.

Dou, Z., Kotekal, S., Xu, Z., and Zhou, H. H. (2024). From optimal score matching to optimal sampling. *arXiv preprint arXiv:2409.07032*.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614.

Eldan, R. (2020). Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation. *Probability Theory and Related Fields*, 176(3):737–755.

Eldan, R. (2022). Analysis of high-dimensional distributions using pathwise methods. In *Proc. Int. Cong. Math*, volume 6, pages 4246–4270.

Feng, O. Y., Kao, Y.-C., Xu, M., and Samworth, R. J. (2024). Optimal convex m-estimation via score matching. *arXiv preprint arXiv:2403.16688*.

Fu, H., Yang, Z., Wang, M., and Chen, M. (2024). Unveil conditional diffusion models with classifier-free guidance: A sharp statistical theory. *arXiv preprint arXiv:2403.11968*.

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

Gentiloni-Silveri, M. and Ocello, A. (2025). Beyond log-concavity and score regularity: Improved convergence bounds for score-based generative models in w2-distance. *arXiv preprint arXiv:2501.02298*.

Gupta, S., Cai, L., and Chen, S. (2024). Faster diffusion-based sampling with randomized midpoints: Sequential and parallel. *arXiv preprint arXiv:2406.00924*.

Han, Y., Razaviyayn, M., and Xu, R. (2024). Neural network-based score estimation in diffusion models: Optimization and generalization. *arXiv preprint arXiv:2401.15604*.

Haussmann, U. G. and Pardoux, E. (1986). Time reversal of diffusions. *The Annals of Probability*, pages 1188–1205.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

Huang, D. Z., Huang, J., and Lin, Z. (2024a). Convergence analysis of probability flow ode for score-based generative models. *arXiv preprint arXiv:2404.09730*.

Huang, Z., Wei, Y., and Chen, Y. (2024b). Denoising diffusion probabilistic models are optimally adaptive to unknown low dimensionality. *arXiv preprint arXiv:2410.18784*.

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

Hyvärinen, A. (2007). Some extensions of score matching. *Computational statistics & data analysis*, 51(5):2499–2512.

Kazerouni, A., Aghdam, E. K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., and Merhof, D. (2023). Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846.

Koehler, F., Heckett, A., and Risteski, A. (2023). Statistical efficiency of score matching: The view from isoperimetry. *International Conference on Learning Representations*.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pages 1302–1338.

Lee, H., Lu, J., and Tan, Y. (2022). Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882.

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

Li, G. and Cai, C. (2024). Provable acceleration for diffusion models under minimal assumptions. *arXiv preprint arXiv:2410.23285*.

Li, G., Huang, Y., Efimov, T., Wei, Y., Chi, Y., and Chen, Y. (2024a). Accelerating convergence of score-based diffusion models, provably. *arXiv preprint arXiv:2403.03852*.

Li, G., Huang, Z., and Wei, Y. (2024b). Towards a mathematical theory for consistency training in diffusion models. *arXiv preprint arXiv:2402.07802*.

Li, G. and Jiao, Y. (2024). Improved convergence rate for diffusion probabilistic models. *arXiv preprint arXiv:2410.13738*.

Li, G., Wei, Y., Chen, Y., and Chi, Y. (2023a). Towards faster non-asymptotic convergence for diffusion-based generative models. *arXiv preprint arXiv:2306.09251*.

Li, G., Wei, Y., Chi, Y., and Chen, Y. (2024c). A sharp convergence theory for the probability flow ODEs of diffusion models. *arXiv preprint arXiv:2408.02320*.

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

Li, G. and Yan, Y. (2024b). $O(d/T)$ convergence theory for diffusion probabilistic models under minimal assumptions. *arXiv preprint arXiv:2409.18959*.

Li, P., Li, Z., Zhang, H., and Bian, J. (2023b). On the generalization properties of diffusion models. *Advances in Neural Information Processing Systems*, 36:2097–2127.

Li, R., Di, Q., and Gu, Q. (2024d). Unified convergence analysis for score-based diffusion models with deterministic samplers. *arXiv preprint arXiv:2410.14237*.

Li, W., Zhang, H., and Qu, Q. (2024e). Shallow diffuse: Robust and invisible watermarking through low-dimensional subspaces in diffusion models. *arXiv preprint arXiv:2410.21088*.

Li, X., Dai, Y., and Qu, Q. (2024f). Understanding generalizability of diffusion models requires rethinking the hidden gaussian structure. *arXiv preprint arXiv:2410.24060*.

Liang, Y., Ju, P., Liang, Y., and Shroff, N. (2024). Broadening target distributions for accelerated diffusion models via a novel analysis approach. *arXiv preprint arXiv:2402.13901*.

Lin, L., Li, Z., Li, R., Li, X., and Gao, J. (2024). Diffusion models for time-series applications: a survey. *Frontiers of Information Technology & Electronic Engineering*, 25(1):19–41.

Liu, X., Wu, L., Ye, M., and Liu, Q. (2022). Let us build bridges: Understanding and extending diffusion generative models. *arXiv preprint arXiv:2208.14699*.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. (2022a). Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. (2022b). Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*.

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

Oko, K., Akiyama, S., and Suzuki, T. (2023). Diffusion models are minimax optimal distribution estimators. *arXiv preprint arXiv:2303.01861*.

Pedrotti, F., Maas, J., and Mondelli, M. (2023). Improved convergence of score-based diffusion models via prediction-correction. *arXiv preprint arXiv:2305.14164*.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. (2021). The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*.

Potaptchik, P., Azangulov, I., and Deligiannidis, G. (2024). Linear convergence of diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2410.09046*.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Ren, Y., Chen, H., Rotskoff, G. M., and Ying, L. (2024). How discrete and continuous diffusion meet: Comprehensive analysis of discrete diffusion models via a stochastic integral framework. *arXiv preprint arXiv:2410.03601*.

Robbins, H. E. (1992). An empirical bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and basic theory*, pages 388–394. Springer.

Shen, R. and Lee, Y. T. (2019). The randomized midpoint method for log-concave sampling. *Advances in Neural Information Processing Systems*, 32.

Song, J., Meng, C., and Ermon, S. (2020). Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y. and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

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

Stanczuk, J. P., Batzolis, G., Deveney, T., and Schönlieb, C.-B. (2024). Diffusion models encode the intrinsic dimension of data manifolds. In *Forty-first International Conference on Machine Learning*.

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

Tang, W. (2023). Diffusion probabilistic models. *preprint*.

Tang, W. and Xu, R. (2024). A stochastic analysis approach to conditional diffusion guidance.

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

Tang, W. and Zhao, H. (2024b). Score-based diffusion models via stochastic differential equations–a technical tutorial. *arXiv preprint arXiv:2402.07487*.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

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

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Wang, P., Zhang, H., Zhang, Z., Chen, S., Ma, Y., and Qu, Q. (2024). Diffusion models learn low-dimensional distributions via subspace clustering. *arXiv preprint arXiv:2409.02426*.

Wibisono, A., Wu, Y., and Yang, K. Y. (2024). Optimal score estimation via empirical bayes smoothing. *arXiv preprint arXiv:2402.07747*.

Wu, Y., Chen, M., Li, Z., Wang, M., and Wei, Y. (2024a). Theoretical insights for diffusion guidance: A case study for gaussian mixture models. *preprint*.

Wu, Y., Chen, Y., and Wei, Y. (2024b). Stochastic runge-kutta methods: Provable acceleration of diffusion models. *arXiv preprint arXiv:2410.04760*.

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

Zhang, K., Yin, C. H., Liang, F., and Liu, J. (2024). Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. *arXiv preprint arXiv:2402.15602*.

A Technical lemmas

In this section, we present several technical lemmas that prove useful for establishing our main theorems, with their proofs deferred to Appendix G. For simplicity of presentation, we assume without loss of generality that $k \geq \log d$ throughout the proof.

Before proceeding, let us introduce several notation that will be useful throughout.

- Let $\{x_i^\star\}_{1\leq i\leq N_{\epsilon_0}}$ be an ϵ_0-net of $\mathcal{X}_{\mathsf{data}}$, with N_{ϵ_0} denote its cardinality. Let $\{\mathcal{B}_i\}_{1\leq i\leq N_{\epsilon_0}}$ be a disjoint ϵ_0-cover for $\mathcal{X}_{\mathsf{data}}$ such that $x_i^\star \in \mathcal{B}_i$ for each i. See, e.g., Vershynin (2018), for the definition of epsilon-net and epsilon-cover.

- Define the following two sets:

$$\mathcal{I} := \{1 \leq i \leq N_{\epsilon_0} : \mathbb{P}(X_0 \in \mathcal{B}_i) \geq \exp(-C_1 k \log T)\} \tag{37}$$

  and

$$\mathcal{G} := \left\{\omega \in \mathbb{R}^d : \|\omega\|_2 \leq 2\sqrt{d} + \sqrt{C_1 k \log T},\ \left|(x_i^\star - x_j^\star)^\top \omega\right| \leq \sqrt{C_1 k \log T}\ \|x_i^\star - x_j^\star\|_2,\ \forall 1 \leq i,j \leq N_{\epsilon_0}\right\}.$$

  for some sufficiently large universal constant $C_1 > 0$. As it turns out, both $\bigcup_{i\in\mathcal{I}} \mathcal{B}_i$ and $\mathcal{G}$ can be utilized to define certain high-probability sets related to the random vector $X_0 \sim p_{\mathsf{data}}$ and a standard Gaussian random vector in $\mathbb{R}^d$.

- Let us express

$$X_t = \sqrt{\overline{\alpha}_t} X_0 + \sqrt{1-\overline{\alpha}_t} Z \tag{38}$$

for some random vector $Z \sim \mathcal{N}(0, I_d)$. By taking

$$V_\alpha \coloneqq \sqrt{\alpha} V_1 + \sqrt{1-\alpha} Z \qquad \text{with } V_1 \coloneqq X_0, \tag{39}$$

we see that

$$X_t = V_{\overline{\alpha}_t} \tag{40}$$

for any t. For every $\alpha \in [0, 1-1/T]$, define a typical set for each V_α as follows

$$\mathcal{T}_\alpha \coloneqq \left\{ \sqrt{\alpha} v_1 + \sqrt{1-\alpha}\, \omega : v_1 \in \bigcup_{i \in \mathcal{I}} \mathcal{B}_i, \omega \in \mathcal{G} \right\}. \tag{41}$$

- Next, we turn to the posterior distribution of V_1 given V_α, which dictates the performance of DDIM and DDPM samplers. Let us introduce the following shorthand notation:

$$\begin{aligned} \mu_{V_1|V_\alpha}(v) &\coloneqq \mathbb{E}[V_1 \mid V_\alpha = v], \\ \mathrm{Cov}_{V_1|V_\alpha}(v) &\coloneqq \mathbb{E}[V_1 V_1^\top \mid V_\alpha = v] - \mu_{V_1|V_\alpha}(v) \mu_{V_1|V_\alpha}(v)^\top. \end{aligned} \tag{42}$$

Given that the random objects $\mu_{V_1|V_{\overline{\alpha}_t}}(V_{\overline{\alpha}_t})$ and $\mathrm{Cov}_{V_1|V_{\overline{\alpha}_t}}(V_{\overline{\alpha}_t})$ will be used frequently throughout the proof, we shall often employ the following shorthand notation

$$\mu_{0|t} \coloneqq \mu_{V_1|V_{\overline{\alpha}_t}}(V_{\overline{\alpha}_t}) \qquad \text{and} \qquad \mathrm{Cov}_{0|t} \coloneqq \mathrm{Cov}_{V_1|V_{\overline{\alpha}_t}}(V_{\overline{\alpha}_t}) \tag{43}$$

as long as it is clear from the context.

- In addition, we find it convenient to define

$$\varepsilon_t^{\mathsf{sc}}(x) \coloneqq s_t(x) - s_t^*(x) \qquad \text{and} \qquad \varepsilon_t^{\mathsf{J}}(x) \coloneqq \frac{\partial s_t(x)}{\partial x} - \frac{\partial s_t^*(x)}{\partial x}. \tag{44}$$

Now, let us proceed to present the technical lemmas. While some of these proofs can be found in Li and Yan (2024a); Huang et al. (2024b), we provide them here for the sake of completeness.

The first lemma demonstrates that, for any $\alpha \in [0, 1-1/T]$, $\mathcal{T}_\alpha$ is a high-probability set for V_α. The proof of this result is deferred to Appendix G.1.

Lemma 1 *There exists some universal constant $C_1 \gg C_{\mathsf{cover}}$ such that for any $\alpha \in [0, 1-1/T]$, we have*

$$\mathbb{P}\left(V_\alpha \notin \mathcal{T}_\alpha\right) \le \exp\left(-\frac{C_1}{4} k \log T\right).$$

Next, we develop a lemma that characterizes the concentration property of the point V_1 given the observation V_α; the proof can be found in Appendix G.2.

Lemma 2 *Consider any $v \in \mathcal{T}_\alpha$, and let us write $v = \sqrt{\alpha} v_1^* + \sqrt{1-\alpha}\, \omega$ for some $v_1^* \in \bigcup_{i \in \mathcal{I}} \mathcal{B}_i$ and $\omega \in \mathcal{G}$ (cf. (41)). Suppose that $v_1^* \in \mathcal{B}_{i(v)}$ for some $i(v) \in \mathcal{I}$. Then there exists some universal constant $C_2 > 0$ such that for any $C \ge C_2$,*

$$\mathbb{P}\left(\sqrt{\alpha} \left\| V_1 - x_{i(v)}^* \right\|_2 \ge \sqrt{Ck(1-\alpha)\log T} \mid V_\alpha = v\right) \le \exp\left(-\frac{C}{20} k \log T\right).$$

Armed with the above lemma, we can readily establish, for any $\alpha \in [0, 1-1/T]$, upper bounds on the moments of V_1 under the posterior distribution $\mathbb{P}(\cdot \mid V_\alpha = v)$, provided that $v \in \mathcal{T}_\alpha$. This is stated in the following corollary.

Corollary 1 *There exists a universal constant $C_3 > 0$, such that for any $\alpha \in [0, 1 - 1/T]$, the following inequalities hold for any point $v \in \mathcal{T}_\alpha$:*

$$\mathbb{E}\left[\left\|V_1 - \mu_{V_1|V_\alpha}(v)\right\|_2^l \mid V_\alpha = v\right] \le C_3 \left(\frac{1-\alpha}{\alpha} k \log T\right)^{l/2}, \ l = 1, 2, 3, 4. \tag{45}$$

Moreover, we single out the lemma below that can help control the posterior covariance of interest. The proof is postponed to Appendix G.3.

Lemma 3 *Suppose that Assumptions 1 and 2 hold. Denote $\widetilde{\sigma}_t^2 = \frac{\overline{\alpha}_t(1-\alpha_t)}{(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)}$. Then for any $t \ge 2$, the posterior covariance defined in (42) satisfies*

$$\widetilde{\sigma}_t^2 \mathbb{E}_{X_t}\left[\left\|\mathrm{Cov}_{0|t}\right\|_{\mathrm{F}}^2\right] \le 3\Big\{\mathbb{E}\big[\mathrm{tr}(\mathrm{Cov}_{0|t})\big] - \mathbb{E}\big[\mathrm{tr}(\mathrm{Cov}_{0|t-1})\big]\Big\} + \frac{3}{T^{10}}.$$

We also make note of the following basic property about $\{\alpha_t\}$ (see Li et al. (2024c, Section 5.1)):

$$\frac{1}{2}\frac{1-\alpha_t}{1-\overline{\alpha}_t} \le \frac{1}{2}\frac{1-\alpha_t}{\alpha_t - \overline{\alpha}_t} \le \frac{1-\alpha_t}{1-\overline{\alpha}_{t-1}} \le \frac{4c_1 \log T}{T} \qquad \text{for any } 2 \le t \le T. \tag{46}$$

The lemma below is a consequence of this property, whose proof can be found in Appendix G.4.

Lemma 4 *There exists some universal constant $C_6 > 0$ such that for any $t \ge 1$,*

$$\frac{\overline{\alpha}_t(1-\alpha_t)}{(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\overline{\alpha}_{t+1}(1-\alpha_{t+1})}{(\alpha_{t+1} - \overline{\alpha}_{t+1})(1-\overline{\alpha}_{t+1})} \le \frac{C_6 \log^2 T}{T^2} \frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}.$$

In addition, Lemma 5 below helps control the tightness of the second-order Taylor expansion of a certain log-determinant function. Its proof is deferred to Appendix G.5.

Lemma 5 *Let $A \in \mathbb{R}^{d\times d}$ be any positive semi-definite matrix, and $\Delta \in \mathbb{R}^{d\times d}$ be any square matrix. Suppose $\eta\|\Delta\| \le \frac{1}{4}$, where $0 < \eta < 1$. Then it holds that*

$$\log\det\left(I + \eta A + \eta\Delta\right) \ge \eta\big(\mathrm{tr}(A) + \mathrm{tr}(\Delta)\big) - 4\eta^2\left(\|A\|_{\mathrm{F}}^2 + \|\Delta\|_{\mathrm{F}}^2\right).$$

Finally, the Tweedie formula as stated below (Efron, 2011), which establishes the intimate connection between the score function (resp. its corresponding Jacobian matrix) and the posterior mean (resp. posterior covariance) of X_0 given X_t, will be invoked multiple times. For its proof, one can refer to Robbins (1992).

$$\begin{aligned} s_t^*(x_t) &= \frac{\sqrt{\overline{\alpha}_t}}{1-\overline{\alpha}_t}\mu_{0|t}(x_t) - \frac{1}{1-\overline{\alpha}_t}x_t, \\ \frac{\partial s_t^*(x_t)}{\partial x_t} &= \frac{\overline{\alpha}_t}{(1-\overline{\alpha}_t)^2}\mathrm{Cov}_{0|t}(x_t) - \frac{1}{1-\overline{\alpha}_t}I. \end{aligned} \tag{47}$$

where

$$\mu_{0|t}(x_t) = \mathbb{E}[X_0 \mid X_t = x_t], \tag{48a}$$

$$\mathrm{Cov}_{0|t}(x_t) = \mathbb{E}[X_0 X_0^\top \mid X_t = x_t] - \mathbb{E}[X_0 \mid X_t = x_t]\,\mathbb{E}[X_0 \mid X_t = x_t]^\top. \tag{48b}$$

B Analysis for DDIM (proof of Theorem 1)

In this section, we establish our convergence guarantees for the DDIM sampler as stated in Theorem 1. Throughout this section, we define the (deterministic) mapping $\Phi_t(\cdot)$ as

$$\Phi_t(x) = \frac{1}{\sqrt{\alpha_t}}\big(x + \eta_t s_t(x)\big) \qquad \text{with } \eta_t = \frac{1-\alpha_t}{1 + \sqrt{\frac{\alpha_t - \overline{\alpha}_t}{1-\overline{\alpha}_t}}}, \tag{49}$$

where $s_t(x)$ is the score estimate. According to the update rule (10) and (11a) of DDIM, we know that (since $\eta_t^{\mathsf{ddim}} = \eta_t$)

$$Y_{t-1} = \Phi_t(Y_t).$$

We begin by isolating the following recursion that plays an important role in our analysis: for any $t \geq 2$,

$$\begin{aligned}
\mathsf{TV}(p_{X_{t-1}}, p_{Y_{t-1}}) &= \sup_{\mathcal{A}\subseteq\mathcal{X}} \left\{\mathbb{P}_{X_{t-1}}(\mathcal{A}) - \mathbb{P}_{Y_{t-1}}(\mathcal{A})\right\} = \sup_{\mathcal{A}\subseteq\mathcal{X}} \left\{\mathbb{P}_{X_{t-1}}(\mathcal{A}) - \mathbb{P}_{Y_t}\big({\Phi_t}^{-1}(\mathcal{A})\big)\right\} \\
&\leq \sup_{\mathcal{A}\subseteq\mathcal{X}} \left\{\mathbb{P}_{X_{t-1}}(\mathcal{A}) - \mathbb{P}_{X_t}\big({\Phi_t}^{-1}(\mathcal{A})\big)\right\} + \sup_{\mathcal{A}\subseteq\mathcal{X}} \left\{\mathbb{P}_{X_t}\big({\Phi_t}^{-1}(\mathcal{A})\big) - \mathbb{P}_{Y_t}\big({\Phi_t}^{-1}(\mathcal{A})\big)\right\} \\
&\leq \sup_{\mathcal{A}} \left\{\mathbb{P}_{X_{t-1}}(\mathcal{A}) - \mathbb{P}_{\Phi_t(X_t)}(\mathcal{A})\right\} + \mathsf{TV}(p_{X_t}, p_{Y_t}) \\
&= \mathsf{TV}(p_{X_{t-1}}, p_{\Phi_t(X_t)}) + \mathsf{TV}(p_{X_t}, p_{Y_t}),
\end{aligned} \tag{50}$$

where the first identity arises from the basic property of the TV distance.

Next, we define a set $\mathcal{E}_t$ as

$$\mathcal{E}_t := \left\{x \in \mathcal{X}_{\mathsf{data}} : \eta_t \left\|\varepsilon_t^{\mathsf{J}}(x)\right\| \leq \frac{1}{8}\right\}. \tag{51}$$

Based on the set $\mathcal{E}_t$, we define an auxiliary map $\widetilde{\Phi}_t : \mathcal{X}_{\mathsf{data}} \to \mathcal{R}^d \cup \{\infty\}$ as

$$\widetilde{\Phi}_t(x) = \begin{cases} \Phi_t(x), & \text{if } x \in \mathcal{E}_t, \\ \infty, & \text{otherwise.} \end{cases} \tag{52}$$

Further, for any vector $v \in \mathbb{R}^d$ with $\|v\|_2 = 1$ and any $x \in \mathcal{E}_t$, Tweedie's formula (47) tells us that

$$\begin{aligned}
v^\top \frac{\partial \widetilde{\Phi}_t(x)}{\partial x} v = v^\top \frac{\partial \Phi_t(x)}{\partial x} v &= v^\top \left(I + \eta_t \frac{\partial s_t(x)}{\partial x}\right) v = v^\top \left(I + \eta_t \frac{\partial s_t^*(x)}{\partial x}\right) v + \eta_t v^\top \varepsilon_t^{\mathsf{J}}(x) v \\
&\geq v^\top \left(I + \eta_t \frac{\partial s_t^*(x)}{\partial x}\right) v - \frac{1}{8}\|v\|_2^2 = v^\top \left\{\left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right) I + \frac{\overline{\alpha}_t \eta_t}{(1-\overline{\alpha}_t)^2} \mathrm{Cov}_{0|t}(x)\right\} v - \frac{1}{8}\|v\|_2^2 \\
&\geq \left(1 - \frac{\eta_t}{1-\overline{\alpha}_t} - \frac{1}{8}\right)\|v\|_2^2 > \frac{3}{8}.
\end{aligned}$$

Here, the penultimate relation follows from the positive semidefiniteness of the covariance matrix $\mathrm{Cov}_{0|t}(x)$; and the last inequality holds since

$$\frac{\eta_t}{1-\overline{\alpha}_t} = \frac{1-\alpha_t}{1-\overline{\alpha}_t + \sqrt{(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)}} \leq \frac{1-\alpha_t}{2(\alpha_t - \overline{\alpha}_t)} \leq \frac{4c_1 \log T}{T} < \frac{1}{4},$$

a consequence of property (46). This result implies that $\frac{\partial \widetilde{\Phi}_t(x)}{\partial x}$ is positive definite uniformly over all $x \in \mathcal{E}_t$. According to the inverse mapping theorem, we know that $\widetilde{\Phi}_t$ is a bijection on $\mathcal{E}_t$. Therefore, for any $x_{t-1} \in \widetilde{\Phi}_t(\mathcal{E}_t)$, we can find a unique $x_t \in \mathcal{E}_t$ such that $\widetilde{\Phi}_t(x_t) = x_{t-1}$, which in turn allows us to derive

$$p_{\widetilde{\Phi}_t(X_t)}(x_{t-1}) = p_{X_t}\big(\widetilde{\Phi}_t^{-1}(x_{t-1})\big) \cdot \det\left(\frac{\partial \widetilde{\Phi}_t^{-1}(x_{t-1})}{\partial x_{t-1}}\right) = p_{X_t}(x_t) \cdot \det\left(\frac{\partial x_t}{\partial x_{t-1}}\right).$$

Consequently, we see that: for any $t \geq 2$ and any $x_{t-1} \in \widetilde{\Phi}_t(\mathcal{E}_t)$,

$$\begin{aligned}
p_{\widetilde{\Phi}_t(X_t)}(x_{t-1}) - p_{X_{t-1}}(x_{t-1}) &= p_{X_t}(x_t) \det\left(\frac{\partial x_t}{\partial x_{t-1}}\right) - p_{X_{t-1}}(x_{t-1}) \\
&= \int \left\{ p_{X_t|X_0}(x_t \mid x_0) \det\left(\frac{\partial x_t}{\partial x_{t-1}}\right) - p_{X_{t-1}|X_0}(x_{t-1} \mid x_0) \right\} p_{X_0}(x_0) \mathrm{d}x_0 \\
&= \int \left\{ 1 - \underbrace{\frac{p_{X_{t-1}|X_0}(x_{t-1} \mid x_0)}{p_{X_t|X_0}(x_t \mid x_0)} \det\left(\frac{\partial x_{t-1}}{\partial x_t}\right)}_{=:\mathcal{T}(x_t, x_0)} \right\} p_{X_t, X_0}(x_t, x_0) \det\left(\frac{\partial x_t}{\partial x_{t-1}}\right) \mathrm{d}x_0 \\
&= \int \{1 - \mathcal{T}(x_t, x_0)\} \, p_{X_t, X_0}(x_t, x_0) \det\left(\frac{\partial x_t}{\partial x_{t-1}}\right) \mathrm{d}x_0.
\end{aligned} \tag{53}$$

The next step is then to analyze $\mathcal{T}(x_t, x_0)$.

Controlling $\mathcal{T}(x_t, x_0)$: Given how X_{t-1} and X_t are generated, we observe that

$$\begin{aligned}
\mathcal{T}(x_t, x_0) =& \frac{p_{X_{t-1}|X_0}(x_{t-1} \mid x_0)}{p_{X_t|X_0}(x_t \mid x_0)} \det\left(\frac{\partial x_{t-1}}{\partial x_t}\right) = \frac{\left(\frac{1}{1-\overline{\alpha}_{t-1}}\right)^{\frac{d}{2}} \exp\left\{\frac{-\|x_{t-1} - \sqrt{\overline{\alpha}_{t-1}} x_0\|_2^2}{2(1-\overline{\alpha}_{t-1})}\right\}}{\left(\frac{1}{1-\overline{\alpha}_t}\right)^{\frac{d}{2}} \exp\left\{\frac{-\|x_t - \sqrt{\overline{\alpha}_t} x_0\|_2^2}{2(1-\overline{\alpha}_t)}\right\}} \det\left(\frac{\partial x_{t-1}}{\partial x_t}\right) \\
=& \underbrace{\left(\frac{1-\overline{\alpha}_t}{1-\overline{\alpha}_{t-1}}\right)^{\frac{d}{2}} \det\left(\frac{\partial x_{t-1}}{\partial x_t}\right)}_{=:\mathcal{T}_1(x_t, x_0)} \underbrace{\exp\left\{\frac{\left\|x_t - \sqrt{\overline{\alpha}_t} x_0\right\|_2^2}{2(1-\overline{\alpha}_t)} - \frac{\|x_{t-1} - \sqrt{\overline{\alpha}_{t-1}} x_0\|_2^2}{2(1-\overline{\alpha}_{t-1})}\right\}}_{=:\mathcal{T}_2(x_t, x_0)},
\end{aligned} \tag{54}$$

leaving us with two terms to control.

Let us first study the term $\mathcal{T}_1(x_t, x_0)$, towards which we see that

$$\begin{aligned}
\mathcal{T}_1(x_t, x_0) &= \left(\frac{1-\overline{\alpha}_t}{1-\overline{\alpha}_{t-1}}\right)^{\frac{d}{2}} \det\left(\frac{\partial (x_t + \eta_t s_t(x_t))/\sqrt{\alpha_t}}{\partial x_t}\right) = \left(\frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}\right)^{\frac{d}{2}} \det\left(\frac{\partial (x_t + \eta_t s_t(x_t))}{\partial x_t}\right) \\
&= \left(\frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}\right)^{\frac{d}{2}} \det\left(I + \eta_t \frac{\partial}{\partial x_t} s_t^*(x_t) + \eta_t \left(\frac{\partial}{\partial x_t} s_t(x_t) - \frac{\partial}{\partial x_t} s_t^*(x_t)\right)\right) \\
&\overset{(a)}{=} \left(\frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}\right)^{\frac{d}{2}} \det\left(I + \eta_t \left\{\frac{\overline{\alpha}_t}{(1-\overline{\alpha}_t)^2} \mathrm{Cov}_{0|t}(x_t) - \frac{1}{1-\overline{\alpha}_t} I\right\} + \eta_t \varepsilon_t^{\mathsf{J}}(x_t)\right) \\
&= \det\left(\sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}} \left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right) I + \sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}} \frac{\overline{\alpha}_t \eta_t}{(1-\overline{\alpha}_t)^2} \mathrm{Cov}_{0|t}(x_t) + \sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}} \eta_t \varepsilon_t^{\mathsf{J}}(x_t)\right) \\
&\overset{(b)}{=} \det\left(I + \sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}} \frac{\overline{\alpha}_t \eta_t}{(1-\overline{\alpha}_t)^2} \mathrm{Cov}_{0|t}(x_t) + \sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}} \eta_t \varepsilon_t^{\mathsf{J}}(x_t)\right).
\end{aligned} \tag{55}$$

Here, (a) arises from Tweedie's formula (47), whereas (b) follows since

$$\sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}} \left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right) = \sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}} \frac{(1-\overline{\alpha}_t) - \eta_t}{1-\overline{\alpha}_t} = \sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}} \frac{\sqrt{(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)}}{1-\overline{\alpha}_t} = 1. \tag{56}$$

Next, we turn attention to the term $\mathcal{T}_2(x_t, x_0)$, which satisfies

$$\begin{aligned}
\log \mathcal{T}_2(x_t, x_0) &= \frac{\left\|x_t - \sqrt{\overline{\alpha}_t} x_0\right\|_2^2}{2(1-\overline{\alpha}_t)} - \frac{\left\|(x_t + \eta_t s_t(x_t))/\sqrt{\alpha_t} - \sqrt{\overline{\alpha}_{t-1}} x_0\right\|_2^2}{2(1-\overline{\alpha}_{t-1})} \\
&= \frac{\left\|x_t - \sqrt{\overline{\alpha}_t} x_0\right\|_2^2}{2(1-\overline{\alpha}_t)} - \frac{\left\|x_t + \eta_t s_t(x_t) - \sqrt{\overline{\alpha}_t} x_0\right\|_2^2}{2(\alpha_t - \overline{\alpha}_t)}.
\end{aligned} \tag{57}$$

Let us first look at the term $x_t + \eta_t s_t(x_t) - \sqrt{\overline{\alpha}_t} x_0$. It follows from Tweedie's formula (47) that

$$\begin{aligned}
x_t + \eta_t s_t(x_t) - \sqrt{\overline{\alpha}_t} x_0 &= \left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right) x_t + \frac{\sqrt{\overline{\alpha}_t}\eta_t}{1-\overline{\alpha}_t}\mu_{0|t}(x_t) - \sqrt{\overline{\alpha}_t} x_0 + \eta_t\big(s_t(x_t) - s_t^*(x_t)\big) \\
&= \left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right)(x_t - \sqrt{\overline{\alpha}_t} x_0) + \frac{\sqrt{\overline{\alpha}_t}\eta_t}{1-\overline{\alpha}_t}\mu_{0|t}(x_t) - \frac{\eta_t}{1-\overline{\alpha}_t}\sqrt{\overline{\alpha}_t} x_0 + \eta_t\big(s_t(x_t) - s_t^*(x_t)\big) \\
&= \left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right)(x_t - \sqrt{\overline{\alpha}_t} x_0) + \frac{\sqrt{\overline{\alpha}_t}\eta_t}{1-\overline{\alpha}_t}\big(\mu_{0|t}(x_t) - x_0\big) + \eta_t\big(s_t(x_t) - s_t^*(x_t)\big) \\
&= \left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right)(x_t - \sqrt{\overline{\alpha}_t} x_0) + \frac{\sqrt{\overline{\alpha}_t}\eta_t}{1-\overline{\alpha}_t}\big(\mu_{0|t}(x_t) - x_0\big) + \eta_t \varepsilon_t^{\mathsf{sc}}(x_t).
\end{aligned}$$

Substitution into (57) yields

$$\begin{aligned}
\log \mathcal{T}_2(x_t, x_0) &= \frac{\left\|x_t - \sqrt{\overline{\alpha}_t} x_0\right\|_2^2}{2(1-\overline{\alpha}_t)} - \frac{1}{2(\alpha_t - \overline{\alpha}_t)}\left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right)^2 \left\|x_t - \sqrt{\overline{\alpha}_t} x_0\right\|_2^2 \\
&+ \left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right)\frac{\sqrt{\overline{\alpha}_t}\eta_t}{(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)}(x_t - \sqrt{\overline{\alpha}_t} x_0)^\top\big(x_0 - \mu_{0|t}(x_t)\big) - \frac{\overline{\alpha}_t \eta_t^2}{2(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)^2}\left\|x_0 - \mu_{0|t}(x_t)\right\|_2^2 \\
&+ \frac{\eta_t}{\alpha_t - \overline{\alpha}_t}\left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right)\big(x_t - \sqrt{\overline{\alpha}_t}\mu_{0|t}(x_t)\big)^\top \varepsilon_t^{\mathsf{sc}}(x_t) + \frac{\eta_t}{\alpha_t - \overline{\alpha}_t}\big(\mu_{0|t}(x_t) - x_0\big)^\top \varepsilon_t^{\mathsf{sc}}(x_t) + \frac{\eta_t^2}{2(\alpha_t - \overline{\alpha}_t)}\left\|\varepsilon_t^{\mathsf{sc}}(x_t)\right\|_2^2.
\end{aligned} \tag{58}$$

We now control each term of the above display. It follows from (56) that

$$\frac{1}{2(1-\overline{\alpha}_t)} - \frac{1}{2(\alpha_t - \overline{\alpha}_t)}\left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right)^2 = \frac{1}{2(1-\overline{\alpha}_t)} - \frac{1}{2(\alpha_t - \overline{\alpha}_t)}\frac{\alpha_t - \overline{\alpha}_t}{1-\overline{\alpha}_t} = 0,$$

thus implying that

$$\frac{\left\|x_t - \sqrt{\overline{\alpha}_t} x_0\right\|_2^2}{2(1-\overline{\alpha}_t)} - \frac{1}{2(\alpha_t - \overline{\alpha}_t)}\left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right)^2 \left\|x_t - \sqrt{\overline{\alpha}_t} x_0\right\|_2^2 = 0.$$

Further, invoke Tweedie's formula (47) once again to reach

$$\frac{\eta_t}{\alpha_t - \overline{\alpha}_t}\left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right)\big(x_t - \sqrt{\overline{\alpha}_t}\mu_{0|t}(x_t)\big)^\top \varepsilon_t^{\mathsf{sc}}(x_t) = -\frac{\eta_t(1 - \overline{\alpha}_t - \eta_t)}{\alpha_t - \overline{\alpha}_t} s_t^*(x_t)^\top \varepsilon_t^{\mathsf{sc}}(x_t).$$

Additionally, consider the following component of $\log \mathcal{T}_2(x_t, x_0)$:

$$\begin{aligned}
&\left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right)\frac{\sqrt{\overline{\alpha}_t}\eta_t}{(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)}(x_t - \sqrt{\overline{\alpha}_t} x_0)^\top\big(x_0 - \mu_{0|t}(x_t)\big) \\
&\qquad - \frac{\overline{\alpha}_t \eta_t^2}{2(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)^2}\left\|x_0 - \mu_{0|t}(x_t)\right\|_2^2 + \frac{\eta_t}{\alpha_t - \overline{\alpha}_t}(\mu_{0|t}(x_t) - x_0)^\top \varepsilon_t^{\mathsf{sc}}(x_t).
\end{aligned}$$

Taking the expectation of the above expression under the conditional distribution $p_{X_0|X_t}$, we find that

$$\begin{aligned}
&-\left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right)\frac{\overline{\alpha}_t \eta_t}{(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)}\mathrm{tr}\big(\mathrm{Cov}_{0|t}(x_t)\big) - \frac{\overline{\alpha}_t \eta_t^2}{2(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)^2}\mathrm{tr}\big(\mathrm{Cov}_{0|t}(x_t)\big) \\
&\qquad = -\left(1 - \frac{\eta_t}{2(1-\overline{\alpha}_t)}\right)\frac{\overline{\alpha}_t \eta_t}{(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)}\mathrm{tr}\big(\mathrm{Cov}_{0|t}(x_t)\big).
\end{aligned}$$

Defining

$$\xi(x_t, x_0) \coloneqq \left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right)\frac{\sqrt{\overline{\alpha}_t}\eta_t}{(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)}(x_t - \sqrt{\overline{\alpha}_t} x_0)^\top\big(x_0 - \mu_{0|t}(x_t)\big) - \frac{\overline{\alpha}_t \eta_t^2}{2(\alpha_t - \overline{\alpha}_t)(1-\overline{\alpha}_t)^2}\left\|x_0 - \mu_{0|t}(x_t)\right\|_2^2$$

$$+\frac{\eta_t}{\alpha_t-\overline{\alpha}_t}\left(\mu_{0|t}(x_t)-x_0\right)^\top\varepsilon_t^{\mathsf{sc}}(x_t)+\left(1-\frac{\eta_t}{2(1-\overline{\alpha}_t)}\right)\frac{\overline{\alpha}_t\eta_t}{(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\mathrm{tr}\big(\mathrm{Cov}_{0|t}(x_t)\big),$$

one can easily verify that

$$\int_{x_0}\xi(x_t,x_0)p_{X_0|X_t}(x_0\mid x_t)=0,\quad\forall x_t\in\mathcal{X}_{\mathsf{data}}.$$

Thus, based on Eqn. (58), we can further simplify $\log\mathcal{T}_2(x_t,x_0)$ as follows:

$$\begin{aligned}\log\mathcal{T}_2(x_t,x_0)&=\xi(x_t,x_0)-\frac{\overline{\alpha}_t\eta_t}{(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\left(1-\frac{\eta_t}{2(1-\overline{\alpha}_t)}\right)\mathrm{tr}\left(\mathrm{Cov}_{0|t}(x_t)\right)\\&\quad+\frac{\eta_t(1-\overline{\alpha}_t-\eta_t)}{\alpha_t-\overline{\alpha}_t}s_t^*(x_t)^\top\varepsilon_t^{\mathsf{sc}}(x_t)-\frac{\eta_t^2}{2(\alpha_t-\overline{\alpha}_t)}\left\|\varepsilon_t^{\mathsf{sc}}(x_t)\right\|_2^2\\&=\xi(x_t,x_0)-\frac{\overline{\alpha}_t\eta_t}{(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\left(1-\frac{\eta_t}{2(1-\overline{\alpha}_t)}\right)\mathrm{tr}\left(\mathrm{Cov}_{0|t}(x_t)\right)\\&\quad+\sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}}\eta_t s_t^*(x_t)^\top\varepsilon_t^{\mathsf{sc}}(x_t)-\frac{\eta_t^2}{2(\alpha_t-\overline{\alpha}_t)}\left\|\varepsilon_t^{\mathsf{sc}}(x_t)\right\|_2^2,\end{aligned}\tag{59}$$

where the last equality holds since

$$\frac{1-\overline{\alpha}_t-\eta_t}{\alpha_t-\overline{\alpha}_t}=\frac{\sqrt{(1-\overline{\alpha}_t)(\alpha_t-\overline{\alpha}_t)}}{\alpha_t-\overline{\alpha}_t}=\sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}}.$$

Now, to streamline the presentation, define

$$\begin{aligned}W(x_t)\coloneqq\log\det&\left(I+\sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}}\frac{\overline{\alpha}_t\eta_t}{(1-\overline{\alpha}_t)^2}\mathrm{Cov}_{0|t}(x_t)+\sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}}\eta_t\varepsilon_t^{\mathsf{J}}(x_t)\right)-\frac{\eta_t^2}{2(\alpha_t-\overline{\alpha}_t)}\left\|\varepsilon_t^{\mathsf{sc}}(x_t)\right\|_2^2\\&-\frac{\overline{\alpha}_t\eta_t}{(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\left(1-\frac{\eta_t}{2(1-\overline{\alpha}_t)}\right)\mathrm{tr}\left(\mathrm{Cov}_{0|t}(x_t)\right)+\sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}}\eta_t s_t^*(x_t)^\top\varepsilon_t^{\mathsf{sc}}(x_t).\end{aligned}$$

Then for any $x_t\in\mathcal{X}_{\mathsf{data}}$, it holds that

$$\int_{x_0}\left(1-e^{\xi(x_t,x_0)}\right)e^{W(x_t)}p_{X_0|X_t}(x_0\mid x_t)\mathrm{d}x_0\le-e^{W(x_t)}\int_{x_0}\xi(x_t,x_0)p_{X_0|X_t}(x_0\mid x_t)\mathrm{d}x_0=0.\tag{60}$$

where the inequality results from the elementary inequality $1-e^x\le-x$ for all $x\in\mathbb{R}$. Putting (53), (54), (55) and (58) together reveals that: for any $\mathcal{A}\subseteq\mathcal{X}_{\mathsf{data}}\cap\mathcal{E}_t$,

$$\begin{aligned}\mathbb{P}_{\widetilde{\Phi}_t(X_t)}(\mathcal{A})-\mathbb{P}_{X_{t-1}}(\mathcal{A})&=\int_{\mathcal{A}}\left\{p_{\widetilde{\Phi}_t(X_t)}(x_{t-1})-p_{X_{t-1}}(x_{t-1})\right\}\mathrm{d}x_{t-1}\\&=\int_{\mathcal{A}\times\mathcal{X}_{\mathsf{data}}}\left\{1-\mathcal{T}(x_t,x_0)\right\}p_{X_t,X_0}(x_t,x_0)\det\left(\frac{\partial x_t}{\partial x_{t-1}}\right)\mathrm{d}x_0\mathrm{d}x_{t-1}\\&=\int_{\widetilde{\Phi}_t^{-1}(\mathcal{A})\times\mathcal{X}_{\mathsf{data}}}\left\{1-\mathcal{T}(x_t,x_0)\right\}p_{X_t,X_0}(x_t,x_0)\mathrm{d}x_0\mathrm{d}x_t\\&=\int_{\widetilde{\Phi}_t^{-1}(\mathcal{A})\times\mathcal{X}_{\mathsf{data}}}\left\{1-e^{\xi(x_t,x_0)}\cdot e^{W(x_t)}\right\}p_{X_t,X_0}(x_t,x_0)\mathrm{d}x_0\mathrm{d}x_t\\&\overset{(a)}{=}\int_{\widetilde{\Phi}_t^{-1}(\mathcal{A})\times\mathcal{X}_{\mathsf{data}}}\left\{\left(1-e^{\xi(x_t,x_0)}\right)e^{W(x_t)}+\left(1-e^{W(x_t)}\right)\right\}p_{X_t,X_0}(x_t,x_0)\mathrm{d}x_0\mathrm{d}x_t\\&\overset{(b)}{\le}\int_{x_t\in\widetilde{\Phi}_t^{-1}(\mathcal{A})}\left\{1-e^{W(x_t)}\right\}p_{X_t}(x_t)\mathrm{d}x_t\le\int_{x_t\in\widetilde{\Phi}_t^{-1}(\mathcal{A})}-W(x_t)p_{X_t}(x_t)\mathrm{d}x_t,\end{aligned}\tag{61}$$

where (a) follows from simple telescoping, (b) invokes (60), and the last inequality follows from the elementary inequality $1-e^x\le-x$.

B.1 Controlling the function $-W(x_t)$

For $x_t \in \mathcal{E}_t$, applying Lemma 5 reveals that

$$\begin{aligned}
&-\log\det\left(I+\sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}}\frac{\overline{\alpha}_t\eta_t}{(1-\overline{\alpha}_t)^2}\mathrm{Cov}_{0|t}(x_t)+\sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}}\eta_t\varepsilon_t^{\mathsf{J}}(x_t)\right)\\
&\quad\le \frac{4\overline{\alpha}_t^2\eta_t^2}{(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)^3}\left\|\mathrm{Cov}_{0|t}(x_t)\right\|_{\mathrm{F}}^2+\frac{4(1-\overline{\alpha}_t)\eta_t^2}{\alpha_t-\overline{\alpha}_t}\left\|\varepsilon_t^{\mathsf{J}}(x_t)\right\|_{\mathrm{F}}^2\\
&\qquad-\sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}}\frac{\overline{\alpha}_t\eta_t}{(1-\overline{\alpha}_t)^2}\mathrm{tr}\left(\mathrm{Cov}_{0|t}(x_t)\right)-\sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}}\eta_t\mathrm{tr}\big(\varepsilon_t^{\mathsf{J}}(x_t)\big),
\end{aligned}$$

provided that $\sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}}\eta_t\left\|\varepsilon_t^{\mathsf{J}}(x_t)\right\|\le\frac{1}{4}$. Combining this with the definition of $W(x_t)$ results in

$$\begin{aligned}
-W(x_t)&=\frac{\overline{\alpha}_t\eta_t}{(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\left(1-\frac{\eta_t}{2(1-\overline{\alpha}_t)}\right)\mathrm{tr}\left(\mathrm{Cov}_{0|t}(x_t)\right)+\frac{\eta_t^2}{2(\alpha_t-\overline{\alpha}_t)}\left\|\varepsilon_t^{\mathsf{sc}}(x_t)\right\|_2^2\\
&\quad-\sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}}\eta_t s_t^*(x_t)^\top\varepsilon_t^{\mathsf{sc}}(x_t)-\log\det\left(I+\sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}}\frac{\overline{\alpha}_t\eta_t}{(1-\overline{\alpha}_t)^2}\mathrm{Cov}_{0|t}(x_t)+\sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}}\eta_t\varepsilon_t^{\mathsf{J}}(x_t)\right)\\
&\le\frac{4\overline{\alpha}_t^2\eta_t^2}{(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)^3}\left\|\mathrm{Cov}_{0|t}(x_t)\right\|_{\mathrm{F}}^2+\frac{4(1-\overline{\alpha}_t)\eta_t^2}{\alpha_t-\overline{\alpha}_t}\left\|\varepsilon_t^{\mathsf{J}}(x_t)\right\|_{\mathrm{F}}^2+\frac{\eta_t^2}{2(\alpha_t-\overline{\alpha}_t)}\left\|\varepsilon_t^{\mathsf{sc}}(x_t)\right\|_2^2\\
&\quad+\left[\frac{\overline{\alpha}_t\eta_t}{(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\left(1-\frac{\eta_t}{2(1-\overline{\alpha}_t)}\right)-\sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}}\frac{\overline{\alpha}_t\eta_t}{(1-\overline{\alpha}_t)^2}\right]\mathrm{tr}\left(\mathrm{Cov}_{0|t}(x_t)\right)\\
&\quad-\sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}}\eta_t\Big(s_t^*(x_t)^\top\varepsilon_t^{\mathsf{sc}}(x_t)+\mathrm{tr}\left(\varepsilon_t^{\mathsf{J}}(x_t)\right)\Big).
\end{aligned}\tag{62}$$

Recalling the choice of the coefficient $\eta_t=(1-\alpha_t)/\big(1+\sqrt{\frac{\alpha_t-\overline{\alpha}_t}{1-\overline{\alpha}_t}}\big)$, we have

$$\begin{aligned}
&\frac{\overline{\alpha}_t\eta_t}{(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\left(1-\frac{\eta_t}{2(1-\overline{\alpha}_t)}\right)-\sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}}\frac{\overline{\alpha}_t\eta_t}{(1-\overline{\alpha}_t)^2}\\
&=\frac{\overline{\alpha}_t\eta_t}{(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\left(1-\frac{\eta_t}{2(1-\overline{\alpha}_t)}-\sqrt{\frac{\alpha_t-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}}\right)\\
&=\frac{\overline{\alpha}_t\eta_t}{(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\left(1-\frac{\eta_t}{1-\overline{\alpha}_t}-\sqrt{\frac{\alpha_t-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}}+\frac{\eta_t}{2(1-\overline{\alpha}_t)}\right)=\frac{\overline{\alpha}_t\eta_t^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)^2},
\end{aligned}$$

where the penultimate equality holds due to Eqn. (56). Substituting this result into (62) yields

$$\begin{aligned}
-W(x_t)&\le\frac{4\overline{\alpha}_t^2\eta_t^2}{(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)^3}\left\|\mathrm{Cov}_{0|t}(x_t)\right\|_{\mathrm{F}}^2+\frac{4(1-\overline{\alpha}_t)\eta_t^2}{\alpha_t-\overline{\alpha}_t}\left\|\varepsilon_t^{\mathsf{J}}(x_t)\right\|_{\mathrm{F}}^2+\frac{\eta_t^2}{2(\alpha_t-\overline{\alpha}_t)}\left\|\varepsilon_t^{\mathsf{sc}}(x_t)\right\|_2^2\\
&\quad+\frac{\overline{\alpha}_t\eta_t^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)^2}\mathrm{tr}\left(\mathrm{Cov}_{0|t}(x_t)\right)-\sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}}\eta_t\underbrace{\Big(s_t^*(x_t)^\top\varepsilon_t^{\mathsf{sc}}(x_t)+\mathrm{tr}\left(\varepsilon_t^{\mathsf{J}}(x_t)\right)\Big)}_{=:\Delta(\varepsilon_t^{\mathsf{sc}}(x_t),\varepsilon_t^{\mathsf{J}}(x_t))}\\
&\le\frac{4\overline{\alpha}_t^2\eta_t^2}{(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)^3}\left\|\mathrm{Cov}_{0|t}(x_t)\right\|_{\mathrm{F}}^2+\frac{4(1-\overline{\alpha}_t)\eta_t^2}{\alpha_t-\overline{\alpha}_t}\left\|\varepsilon_t^{\mathsf{J}}(x_t)\right\|_{\mathrm{F}}^2+\frac{\eta_t^2}{2(\alpha_t-\overline{\alpha}_t)}\left\|\varepsilon_t^{\mathsf{sc}}(x_t)\right\|_2^2\\
&\quad+\frac{\overline{\alpha}_t\eta_t^2}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)^2}\mathrm{tr}\left(\mathrm{Cov}_{0|t}(x_t)\right)-\sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}}\eta_t\Delta\big(\varepsilon_t^{\mathsf{sc}}(x_t),\varepsilon_t^{\mathsf{J}}(x_t)\big).
\end{aligned}\tag{63}$$

B.2 Controlling the term related to score estimation errors

In this subsection, we would like to control the term $\Delta\big(\varepsilon_t^{\mathsf{sc}}(x_t),\varepsilon_t^{\mathsf{J}}(x_t)\big)$ related to the score estimation error. Specifically, we aim to prove the following lemma.

Lemma 6 *For any set $\mathcal{A} \subseteq \mathcal{X}$, we have*

$$\int_{x_t \in \mathcal{A}} -\Delta\big(\varepsilon_t^{\mathsf{sc}}(x_t), \varepsilon_t^{\mathsf{J}}(x_t)\big) p_{X_t}(x_t) \mathrm{d}x_t \le \frac{2}{\sqrt{1-\overline{\alpha}_t}} \left\{ \varepsilon_{\mathsf{score},t} + \varepsilon_{\mathsf{Jacobi},1,t} + \varepsilon_{\mathsf{Jacobi},2,t} + \varepsilon_{\mathsf{Hess},t} \right\}.$$

Proof of Lemma 6. To begin with, consider the inner product term $s_t^*(x_t)^\top \varepsilon_t^{\mathsf{sc}}(x_t)$. For any set $\mathcal{A}$,

$$\begin{aligned}
\int_{x_t \in \mathcal{A}} -s_t^*(x_t)^\top \varepsilon_t^{\mathsf{sc}}(x_t) p_{X_t}(x_t) \mathrm{d}x_t &= \int_{x_t \in \mathcal{A}} \frac{1}{1-\overline{\alpha}_t} \big(x_t - \sqrt{\overline{\alpha}_t}\mu_{0|t}(x_t)\big)^\top \varepsilon_t^{\mathsf{sc}}(x_t) p_{X_t}(x_t) \mathrm{d}x_t \\
&= \int_{\mathcal{A} \times \mathcal{X}_{\mathsf{data}}} \frac{1}{1-\overline{\alpha}_t} (x_t - \sqrt{\overline{\alpha}_t} x_0)^\top \varepsilon_t^{\mathsf{sc}}(x_t) p_{X_t}(x_t) p_{X_0|X_t}(x_0 \mid x_t) \mathrm{d}x_t \mathrm{d}x_0 \\
&= \int_{x_0, x_t \in \mathcal{X}_{\mathsf{data}} \times \mathcal{A}} \frac{1}{1-\overline{\alpha}_t} (x_t - \sqrt{\overline{\alpha}_t} x_0)^\top \varepsilon_t^{\mathsf{sc}}(x_t) p_{X_t|X_0}(x_t \mid x_0) p_{X_0}(x_0) \mathrm{d}x_t \mathrm{d}x_0 \\
&\le \int_{\mathcal{X}_{\mathsf{data}} \times \mathcal{X}_{\mathsf{data}}} \left| \frac{1}{1-\overline{\alpha}_t} (x_t - \sqrt{\overline{\alpha}_t} x_0)^\top \varepsilon_t^{\mathsf{sc}}(x_t) \right| p_{X_t|X_0}(x_t \mid x_0) p_{X_0}(x_0) \mathrm{d}x_t \mathrm{d}x_0.
\end{aligned} \tag{64}$$

For any given point $x_0 \in \mathcal{X}_{\mathsf{data}}$, applying the Cauchy-Schwartz inequality gives

$$\begin{aligned}
&\int_{\mathcal{X}_{\mathsf{data}}} \left| \frac{1}{1-\overline{\alpha}_t} (x_t - \sqrt{\overline{\alpha}_t} x_0)^\top \varepsilon_t^{\mathsf{sc}}(x_t) \right| p_{X_t|X_0}(x_t \mid x_0) \mathrm{d}x_t \\
&\le \left(\int_{\mathcal{X}_{\mathsf{data}}} \left(\frac{1}{1-\overline{\alpha}_t} (x_t - \sqrt{\overline{\alpha}_t} x_0)^\top \varepsilon_t^{\mathsf{sc}}(x_t) \right)^2 p_{X_t|X_0}(x_t \mid x_0) \mathrm{d}x_t \right)^{\frac{1}{2}} \\
&\le \left(\int_{\mathcal{X}_{\mathsf{data}}} \frac{1}{(1-\overline{\alpha}_t)^2} \left\langle \varepsilon_t^{\mathsf{sc}}(x_t) \varepsilon_t^{\mathsf{sc}}(x_t)^\top, (x_t - \sqrt{\overline{\alpha}_t} x_0)(x_t - \sqrt{\overline{\alpha}_t} x_0)^\top \right\rangle p_{X_t|X_0}(x_t \mid x_0) \mathrm{d}x_t \right)^{\frac{1}{2}}.
\end{aligned} \tag{65}$$

Note that for given x_0, one has $X_t \mid X_0 = x_0 \sim \mathcal{N}\big(\sqrt{\overline{\alpha}_t} x_0, (1-\overline{\alpha}_t)^{-1} I\big)$. As a result,

$$\begin{aligned}
\nabla_{x_t}^2 p_{X_t|X_0}(x_t \mid x_0) &= \nabla_{x_t}^2 \left\{ \left(\frac{1}{2\pi(1-\overline{\alpha}_t)} \right)^{\frac{d}{2}} \exp\left(-\frac{\left\| x_t - \sqrt{\overline{\alpha}_t} x_0 \right\|_2^2}{2(1-\overline{\alpha}_t)} \right) \right\} \\
&= \left(\frac{1}{2\pi(1-\overline{\alpha}_t)} \right)^{\frac{d}{2}} \nabla_{x_t} \left\{ -\frac{x_t - \sqrt{\overline{\alpha}_t} x_0}{1-\overline{\alpha}_t} \exp\left(-\frac{\left\| x_t - \sqrt{\overline{\alpha}_t} x_0 \right\|_2^2}{2(1-\overline{\alpha}_t)} \right) \right\} \\
&= \left(\frac{1}{2\pi(1-\overline{\alpha}_t)} \right)^{\frac{d}{2}} e^{-\frac{\|x_t - \sqrt{\overline{\alpha}_t} x_0\|_2^2}{2(1-\overline{\alpha}_t)}} \left\{ \frac{(x_t - \sqrt{\overline{\alpha}_t} x_0)(x_t - \sqrt{\overline{\alpha}_t} x_0)^\top}{(1-\overline{\alpha}_t)^2} - \frac{1}{1-\overline{\alpha}_t} I \right\} \\
&= p_{X_t|X_0}(x_t \mid x_0) \left\{ \frac{(x_t - \sqrt{\overline{\alpha}_t} x_0)(x_t - \sqrt{\overline{\alpha}_t} x_0)^\top}{(1-\overline{\alpha}_t)^2} - \frac{1}{1-\overline{\alpha}_t} I \right\}.
\end{aligned}$$

From this, we can derive

$$\frac{(x_t - \sqrt{\overline{\alpha}_t} x_0)(x_t - \sqrt{\overline{\alpha}_t} x_0)^\top}{(1-\overline{\alpha}_t)^2} p_{X_t|X_0}(x_t \mid x_0) = \nabla_{x_t}^2 p_{X_t|X_0}(x_t \mid x_0) + p_{X_t|X_0}(x_t \mid x_0) \frac{1}{1-\overline{\alpha}_t} I.$$

Substituting this into (65) yields

$$\begin{aligned}
&\int_{\mathcal{X}_{\mathsf{data}}} \left| \frac{1}{1-\overline{\alpha}_t} (x_t - \sqrt{\overline{\alpha}_t} x_0)^\top \varepsilon_t^{\mathsf{sc}}(x_t) \right| p_{X_t|X_0}(x_t \mid x_0) \mathrm{d}x_t \\
&\le \left(\int_{\mathcal{X}_{\mathsf{data}}} \left\langle \varepsilon_t^{\mathsf{sc}}(x_t) \varepsilon_t^{\mathsf{sc}}(x_t)^\top, \nabla_{x_t}^2 p_{X_t|X_0}(x_t \mid x_0) + \frac{1}{(1-\overline{\alpha}_t)} p_{X_t|X_0}(x_t \mid x_0) I \right\rangle \mathrm{d}x_t \right)^{\frac{1}{2}} \\
&\le \left(\int_{\mathcal{X}_{\mathsf{data}}} \left\langle \varepsilon_t^{\mathsf{sc}}(x_t) \varepsilon_t^{\mathsf{sc}}(x_t)^\top, \nabla_{x_t}^2 p_{X_t|X_0}(x_t \mid x_0) \right\rangle \mathrm{d}x_t \right)^{\frac{1}{2}} + \frac{1}{\sqrt{1-\overline{\alpha}_t}} \left(\int_{\mathcal{X}_{\mathsf{data}}} \|\varepsilon_t^{\mathsf{sc}}(x_t)\|_2^2 p_{X_t|X_0}(x_t \mid x_0) \mathrm{d}x_t \right)^{\frac{1}{2}},
\end{aligned} \tag{66}$$

where the last inequality follows since $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b},\ \forall a, b \geq 0$. With regards to the first term of the above bound, note that

$$\left\langle \varepsilon_t^{\mathsf{sc}}(x_t)\varepsilon_t^{\mathsf{sc}}(x_t)^\top, \nabla^2 p_{X_t|X_0}(x_t \mid x_0)\right\rangle = \sum_{i=1}^d \sum_{j=1}^d [\varepsilon_t^{\mathsf{sc}}(x_t)]_i [\varepsilon_t^{\mathsf{sc}}(x_t)]_j \frac{\partial^2}{\partial x_i \partial x_j} p_{X_t|X_0}(x_t \mid x_0),$$

where $[v]_i$ represents the i-coordinate of the vector v. We can start by analyzing each (i, j) component. In fact, for any $1 \leq i, j \leq d$, we have

$$\begin{aligned}\int_{\mathcal{X}_{\mathsf{data}}} [\varepsilon_t^{\mathsf{sc}}(x_t)]_i [\varepsilon_t^{\mathsf{sc}}(x_t)]_j \frac{\partial^2}{\partial x_i \partial x_j} p_{X_t|X_0}(x_t \mid x_0) \mathrm{d}x_t &\overset{(a)}{=} - \int_{\mathcal{X}_{\mathsf{data}}} \frac{\partial}{\partial x_i} \left([\varepsilon_t^{\mathsf{sc}}(x_t)]_i [\varepsilon_t^{\mathsf{sc}}(x_t)]_j\right) \frac{\partial}{\partial x_j} p_{X_t|X_0}(x_t \mid x_0) \mathrm{d}x_t \\ &\overset{(b)}{=} \int_{\mathcal{X}_{\mathsf{data}}} \frac{\partial^2}{\partial x_i \partial x_j} \left([\varepsilon_t^{\mathsf{sc}}(x_t)]_i [\varepsilon_t^{\mathsf{sc}}(x_t)]_j\right) p_{X_t|X_0}(x_t \mid x_0) \mathrm{d}x_t,\end{aligned} \tag{67}$$

where (a) applies the integration by parts formula with respect to x_i, and (b) applies integration by parts with respect to x_j. Denoting by $[A]_{ij}$ the (i, j)-th element of the matrix A, we have

$$\begin{aligned}\frac{\partial^2}{\partial x_i \partial x_j} \left([\varepsilon_t^{\mathsf{sc}}(x_t)]_i [\varepsilon_t^{\mathsf{sc}}(x_t)]_j\right) = &\left(\frac{\partial^2}{\partial x_i \partial x_j} [\varepsilon_t^{\mathsf{sc}}(x_t)]_i\right) [\varepsilon_t^{\mathsf{sc}}(x_t)]_j + [\varepsilon_t^{\mathsf{sc}}(x_t)]_i \left(\frac{\partial^2}{\partial x_i \partial x_j} [\varepsilon_t^{\mathsf{sc}}(x_t)]_j\right) \\ &+ [\varepsilon_t^{\mathsf{J}}(x_t)]_{ij} [\varepsilon_t^{\mathsf{J}}(x_t)]_{ji} + [\varepsilon_t^{\mathsf{J}}(x_t)]_{ii} [\varepsilon_t^{\mathsf{J}}(x_t)]_{jj}.\end{aligned}$$

Substitution into (67) yields

$$\begin{aligned}&\int_{\mathcal{X}_{\mathsf{data}}} \left\langle \varepsilon_t^{\mathsf{sc}}(x_t)\varepsilon_t^{\mathsf{sc}}(x_t)^\top, \nabla^2 p_{X_t|X_0}(x_t \mid x_0)\right\rangle \mathrm{d}x_t = \sum_{i=1}^d \sum_{j=1}^d \int_{\mathcal{X}_{\mathsf{data}}} [\varepsilon_t^{\mathsf{sc}}(x_t)]_i [\varepsilon_t^{\mathsf{sc}}(x_t)]_j \frac{\partial^2}{\partial x_i \partial x_j} p_{X_t|X_0}(x_t \mid x_0) \mathrm{d}x_t \\ &= \sum_{i=1}^d \sum_{j=1}^d \int_{\mathcal{X}_{\mathsf{data}}} \frac{\partial^2}{\partial x_i \partial x_j} \left([\varepsilon_t^{\mathsf{sc}}(x_t)]_i [\varepsilon_t^{\mathsf{sc}}(x_t)]_j\right) p_{X_t|X_0}(x_t \mid x_0) \mathrm{d}x_t \\ &= \sum_{i=1}^d \sum_{j=1}^d \int_{\mathcal{X}_{\mathsf{data}}} \left\{\left(\frac{\partial^2}{\partial x_i \partial x_j} [\varepsilon_t^{\mathsf{sc}}(x_t)]_i\right) [\varepsilon_t^{\mathsf{sc}}(x_t)]_j + [\varepsilon_t^{\mathsf{sc}}(x_t)]_i \left(\frac{\partial^2}{\partial x_i \partial x_j} [\varepsilon_t^{\mathsf{sc}}(x_t)]_j\right)\right\} p_{X_t|X_0}(x_t \mid x_0) \mathrm{d}x_t \\ &\quad + \sum_{i=1}^d \sum_{j=1}^d \int_{\mathcal{X}_{\mathsf{data}}} \left\{[\varepsilon_t^{\mathsf{J}}(x_t)]_{ij} [\varepsilon_t^{\mathsf{J}}(x_t)]_{ji} + [\varepsilon_t^{\mathsf{J}}(x_t)]_{ii} [\varepsilon_t^{\mathsf{J}}(x_t)]_{jj}\right\} p_{X_t|X_0}(x_t \mid x_0) \mathrm{d}x_t.\end{aligned} \tag{68}$$

We now proceed to investigate each term in the above expression. For any two indices i and j, we observe that

$$\left(\frac{\partial^2}{\partial x_i \partial x_j} [\varepsilon_t^{\mathsf{sc}}(x_t)]_i\right) [\varepsilon_t^{\mathsf{sc}}(x_t)]_j = \left(\frac{\partial}{\partial x_j} \left[\varepsilon_t^{\mathsf{J}}(x_t)\right]_{ii}\right) [\varepsilon_t^{\mathsf{sc}}(x_t)]_j.$$

Consequently, we have

$$\begin{aligned}&\sum_{i=1}^d \sum_{j=1}^d \int_{\mathcal{X}_{\mathsf{data}}} \left\{\left(\frac{\partial^2}{\partial x_i \partial x_j} [\varepsilon_t^{\mathsf{sc}}(x_t)]_i\right) [\varepsilon_t^{\mathsf{sc}}(x_t)]_j + [\varepsilon_t^{\mathsf{sc}}(x_t)]_i \left(\frac{\partial^2}{\partial x_i \partial x_j} [\varepsilon_t^{\mathsf{sc}}(x_t)]_j\right)\right\} p_{X_t|X_0}(x_t \mid x_0) \mathrm{d}x_t \\ &\qquad = 2 \sum_{j=1}^d \int_{\mathcal{X}_{\mathsf{data}}} \frac{\partial}{\partial x_j} \left(\sum_{i=1}^d [\varepsilon_t^{\mathsf{J}}(x_t)]_{ii}\right) [\varepsilon_t^{\mathsf{sc}}(x_t)]_j p_{X_t|X_0}(x_t \mid x_0) \mathrm{d}x_t \\ &\qquad = \int_{\mathcal{X}_{\mathsf{data}}} 2 \left\langle \nabla \mathrm{tr}\left(\varepsilon_t^{\mathsf{J}}(x_t)\right), \varepsilon_t^{\mathsf{sc}}(x_t)\right\rangle p_{X_t|X_0}(x_t \mid x_0) \mathrm{d}x_t\end{aligned}$$

Next, for terms of the form $[\varepsilon_t^{\mathsf{J}}(x_t)]_{ij}[\varepsilon_t^{\mathsf{J}}(x_t)]_{ji}$ and $[\varepsilon_t^{\mathsf{J}}(x_t)]_{ii}[\varepsilon_t^{\mathsf{J}}(x_t)]_{jj}$, simple calculations yield

$$\sum_{i=1}^{d}\sum_{j=1}^{d}[\varepsilon_t^{\mathsf{J}}(x_t)]_{ij}[\varepsilon_t^{\mathsf{J}}(x_t)]_{ji} = \left\langle \varepsilon_t^{\mathsf{J}}(x_t), \varepsilon_t^{\mathsf{J}}(x_t)\right\rangle = \left\|\varepsilon_t^{\mathsf{J}}(x_t)\right\|_{\mathrm{F}}^2;$$
$$\sum_{i=1}^{d}\sum_{j=1}^{d}[\varepsilon_t^{\mathsf{J}}(x_t)]_{ii}[\varepsilon_t^{\mathsf{J}}(x_t)]_{jj} = \left(\sum_{i=1}^{d}[\varepsilon_t^{\mathsf{J}}(x_t)]_{ii}\right)\left(\sum_{j=1}^{d}[\varepsilon_t^{\mathsf{J}}(x_t)]_{jj}\right) = \mathrm{tr}\big(\varepsilon_t^{\mathsf{J}}(x_t)\big)^2.$$

Substituting these results into (68), we obtain

$$\begin{aligned}
&\int_{\mathcal{X}_{\mathsf{data}}} \left\langle \varepsilon_t^{\mathsf{sc}}(x_t)\varepsilon_t^{\mathsf{sc}}(x_t)^\top, \nabla^2 p_{X_t|X_0}(x_t \mid x_0)\right\rangle \mathrm{d}x_t \\
&= \int_{\mathcal{X}_{\mathsf{data}}} \left\{2\left\langle \nabla \mathrm{tr}\left(\varepsilon_t^{\mathsf{J}}(x_t)\right), \varepsilon_t^{\mathsf{sc}}(x_t)\right\rangle + \left(\mathrm{tr}(\varepsilon_t^{\mathsf{J}}(x_t))\right)^2 + \left\|\varepsilon_t^{\mathsf{J}}(x_t)\right\|_{\mathrm{F}}^2\right\} p_{X_t|X_0}(x_t \mid x_0)\mathrm{d}x_t \\
&\le \int_{\mathcal{X}_{\mathsf{data}}} \left\{\left\|\nabla \mathrm{tr}(\varepsilon_t^{\mathsf{J}}(x_t))\right\|_2^2 + \|\varepsilon_t^{\mathsf{sc}}(x_t)\|_2^2 + \left(\mathrm{tr}(\varepsilon_t^{\mathsf{J}}(x_t))\right)^2 + \left\|\varepsilon_t^{\mathsf{J}}(x_t)\right\|_{\mathrm{F}}^2\right\} p_{X_t|X_0}(x_t \mid x_0)\mathrm{d}x_t.
\end{aligned}$$

Combining the above inequality with (64) and (66), we can obtain

$$\begin{aligned}
\int_{x_t\in\mathcal{A}} -s_t^*(x_t)^\top \varepsilon_t^{\mathsf{sc}}(x_t) p_{X_t}(x_t)\mathrm{d}x_t &\le \frac{1}{\sqrt{1-\overline{\alpha}_t}}\int_{\mathcal{X}_{\mathsf{data}}}\left(\int_{\mathcal{X}_{\mathsf{data}}} \|\varepsilon_t^{\mathsf{sc}}(x_t)\|_2^2 p_{X_t|X_0}(x_t \mid x_0)\mathrm{d}x_t\right)^{\frac{1}{2}} p_{X_0}(x_0)\mathrm{d}x_0 \\
&\quad + \int_{x_0\in\mathcal{X}_{\mathsf{data}}}\left(\int_{\mathcal{X}_{\mathsf{data}}} \left\langle \varepsilon_t^{\mathsf{sc}}(x_t)\varepsilon_t^{\mathsf{sc}}(x_t)^\top, \nabla^2 p_{X_t|X_0}(x_t \mid x_0)\right\rangle \mathrm{d}x_t\right)^{\frac{1}{2}} p_{X_0}(x_0)\mathrm{d}x_0 \\
&\le \left(1+\frac{1}{\sqrt{1-\overline{\alpha}_t}}\right)\left(\int_{\mathcal{X}_{\mathsf{data}}} \|\varepsilon_t^{\mathsf{sc}}(x_t)\|_2^2 p_{X_t}(x_t)\mathrm{d}x_t\right)^{\frac{1}{2}} + \left(\int_{\mathcal{X}_{\mathsf{data}}} \left\|\nabla \mathrm{tr}(\varepsilon_t^{\mathsf{J}}(x_t))\right\|_2^2 p_{X_t}(x_t)\mathrm{d}x_t\right)^{\frac{1}{2}} \\
&\quad + \left(\int_{\mathcal{X}_{\mathsf{data}}} \left\|\varepsilon_t^{\mathsf{J}}(x_t)\right\|_{\mathrm{F}}^2 p_{X_t}(x_t)\mathrm{d}x_t\right)^{\frac{1}{2}} + \left(\int_{\mathcal{X}_{\mathsf{data}}} \mathrm{tr}(\varepsilon_t^{\mathsf{J}}(x_t))^2 p_{X_t}(x_t)\mathrm{d}x_t\right)^{\frac{1}{2}} \\
&\le \frac{2}{\sqrt{1-\overline{\alpha}_t}}\varepsilon_{\mathsf{score},t} + \varepsilon_{\mathsf{Jacobi},1,t} + \varepsilon_{\mathsf{Jacobi},2,t} + \varepsilon_{\mathsf{Hess},t}.
\end{aligned} \tag{69}$$

Thus, we arrive at

$$\begin{aligned}
\int_{x_t\in\mathcal{A}} -\Delta\big(\varepsilon_t^{\mathsf{sc}}(x_t), \varepsilon_t^{\mathsf{J}}(x_t)\big) p_{X_t}(x_t)\mathrm{d}x_t &= \int_{x_t\in\mathcal{A}} \left\{-s_t^*(x_t)^\top \varepsilon_t^{\mathsf{sc}}(x_t) - \mathrm{tr}(\varepsilon_t^{\mathsf{J}}(x_t))\right\} p_{X_t}(x_t)\mathrm{d}x_t \\
&\le \frac{2}{\sqrt{1-\overline{\alpha}_t}}\varepsilon_{\mathsf{score},t} + \varepsilon_{\mathsf{Jacobi},1,t} + \varepsilon_{\mathsf{Jacobi},2,t} + \varepsilon_{\mathsf{Hess},t} + \int_{\mathcal{X}_{\mathsf{data}}} \left|\mathrm{tr}(\varepsilon_t^{\mathsf{J}}(x_t))\right| p_{X_t}(x_t)\mathrm{d}x_t \\
&\le \frac{2}{\sqrt{1-\overline{\alpha}_t}}\left(\varepsilon_{\mathsf{score},t} + \varepsilon_{\mathsf{Jacobi},1,t} + \varepsilon_{\mathsf{Jacobi},2,t} + \varepsilon_{\mathsf{Hess},t}\right).
\end{aligned}$$

B.3 Controlling the TV distance between X_{t-1} and $\Phi_t(X_t)$

In Appendices B.1 and B.2, we have developed bounds for all terms in $-W(x_t)$. In this subsection, we combine these results and substitute them into (61), in order to establish a bound on the TV distance between X_{t-1} and $\Phi_t(X_t)$. Recall the definition of the set $\mathcal{E}_t$ (51) and the function $\widetilde{\Phi}_t$ (52). Then for any set $\mathcal{A} \subseteq \mathcal{X}_{\mathsf{data}}$, we can partition the points in $\mathcal{X}_{\mathsf{data}}$ based on whether they belong to the set $\mathcal{E}_t$ and obtain

$$\begin{aligned}
\mathbb{P}_{\Phi_t(X_t)}(\mathcal{A}) - \mathbb{P}_{X_{t-1}}(\mathcal{A}) &\le \mathbb{P}\big(\Phi_t(X_t)\in\mathcal{A};\ X_t\in\mathcal{E}_t\big) - \mathbb{P}(X_{t-1}\in\mathcal{A}) + \mathbb{P}(X_t\notin\mathcal{E}_t) \\
&\overset{(a)}{=} \mathbb{P}\left(\widetilde{\Phi}_t(X_t)\in\mathcal{A},\ \widetilde{\Phi}_t(X_t)\in\Phi_t(\mathcal{E}_t)\right) - \mathbb{P}_{X_{t-1}}(\mathcal{A}) + \mathbb{P}(X_t\notin\mathcal{E}_t)
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{P}_{\widetilde{\Phi}_t(X_t)}\big(\mathcal{A}\cap\Phi_t(\mathcal{E}_t)\big) - \mathbb{P}_{X_{t-1}}(\mathcal{A}\cap\Phi_t(\mathcal{E}_t)) + \mathbb{P}(X_t\notin\mathcal{E}_t)\\
&\overset{(b)}{\leq} \int_{x_t\in\Phi_t^{-1}(\mathcal{A})\cap\mathcal{E}_t}\left(1-e^{W(x_t)}\right)p_{X_t}(x_t)\mathrm{d}x_t + \mathbb{P}(X_t\notin\mathcal{E}_t).
\end{aligned}
$$

Here, (a) follows since $\Phi_t\equiv\widetilde{\Phi}_t$ on $\mathcal{E}_t$, and (b) holds by combining (61) and the following fact:

$$
\widetilde{\Phi}_t^{-1}\big(\mathcal{A}\cap\Phi_t(\mathcal{E}_t)\big) = \widetilde{\Phi}_t^{-1}(\mathcal{A})\cap\widetilde{\Phi}_t^{-1}\circ\Phi_t(\mathcal{E}_t) = \widetilde{\Phi}_t^{-1}(\mathcal{A})\cap\mathcal{E}_t = \Phi_t^{-1}(\mathcal{A})\cap\mathcal{E}_t.
$$

Next, let us look at the two terms on the right-hand side of the above inequality.

In what follows, we would like to make use of inequality (63), which provides a bound for $-W(x_t)$. Before applying this result, we look at the key coefficients in the inequality. According to Lemma 3, we define $\widetilde{\sigma}_t^2 := \frac{\overline{\alpha}_t(1-\alpha_t)}{(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}$. For the DDIM coefficient choice (11a), it holds that $\eta_t = \frac{1-\alpha_t}{1+\sqrt{\frac{\alpha_t-\overline{\alpha}_t}{1-\overline{\alpha}_t}}} \leq 1-\alpha_t$. Taking these results collectively, we arrive at

$$
\begin{aligned}
\frac{\overline{\alpha}_t^2\eta_t^2}{(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)^3} &\leq \frac{\alpha_t-\overline{\alpha}_t}{1-\overline{\alpha}_t}\cdot\left(\frac{\overline{\alpha}_t(1-\alpha_t)}{(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\right)^2 = \frac{\alpha_t-\overline{\alpha}_t}{1-\overline{\alpha}_t}\widetilde{\sigma}_t^4;\\
\frac{\overline{\alpha}_t\eta_t^2}{(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)^2} &\leq \frac{\eta_t}{1-\overline{\alpha}_t}\cdot\frac{\overline{\alpha}_t(1-\alpha_t)}{(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} = \frac{\widetilde{\sigma}_t^2\eta_t}{1-\overline{\alpha}_t}.
\end{aligned}
$$

Substitution into (63) allows one to control the term $\int_{x_t\in\Phi_t^{-1}(\mathcal{A})\cap\mathcal{E}_t}(-W(x_t))p_{X_t}(x_t)\mathrm{d}x_t$ as follows:

$$
\begin{aligned}
\int_{\Phi_t^{-1}(\mathcal{A})\cap\mathcal{E}_t} -W(x_t)p_{X_t}(x_t)\mathrm{d}x_t &\leq \int_{\Phi_t^{-1}(\mathcal{A})\cap\mathcal{E}_t}\left\{\frac{\widetilde{\sigma}_t^2\eta_t}{2(1-\overline{\alpha}_t)}\mathrm{tr}\left(\mathrm{Cov}_{0|t}(x_t)\right) + \frac{\alpha_t-\overline{\alpha}_t}{1-\overline{\alpha}_t}\widetilde{\sigma}_t^4\left\|\mathrm{Cov}_{0|t}(x_t)\right\|_{\mathrm{F}}^2\right\}p_{X_t}(x_t)\mathrm{d}x_t\\
&\quad+\int_{\Phi_t^{-1}(\mathcal{A})\cap\mathcal{E}_t}\left\{\frac{4(1-\overline{\alpha}_t)\eta_t^2}{\alpha_t-\overline{\alpha}_t}\left\|\varepsilon_t^{\mathsf{J}}(x_t)\right\|_{\mathrm{F}}^2 + \frac{\eta_t^2}{2(\alpha_t-\overline{\alpha}_t)}\left\|\varepsilon_t^{\mathsf{sc}}(x_t)\right\|_2^2\right\}p_{X_t}(x_t)\mathrm{d}x_t\\
&\quad-\sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}}\eta_t\int_{\Phi_t^{-1}(\mathcal{A})\cap\mathcal{E}_t}\Delta\big(\varepsilon_t^{\mathsf{sc}}(x_t),\varepsilon_t^{\mathsf{J}}(x_t)\big)p_{X_t}(x_t)\mathrm{d}x_t\\
&\leq \int_{\mathcal{X}_{\mathsf{data}}}\left\{\frac{\widetilde{\sigma}_t^2\eta_t}{2(1-\overline{\alpha}_t)}\mathrm{tr}\left(\mathrm{Cov}_{0|t}(x_t)\right) + \frac{\alpha_t-\overline{\alpha}_t}{1-\overline{\alpha}_t}\widetilde{\sigma}_t^4\left\|\mathrm{Cov}_{0|t}(x_t)\right\|_{\mathrm{F}}^2\right\}p_{X_t}(x_t)\mathrm{d}x_t\\
&\quad+\int_{\mathcal{X}_{\mathsf{data}}}\left\{\frac{4(1-\overline{\alpha}_t)\eta_t^2}{\alpha_t-\overline{\alpha}_t}\left\|\varepsilon_t^{\mathsf{J}}(x_t)\right\|_{\mathrm{F}}^2 + \frac{\eta_t^2}{2(\alpha_t-\overline{\alpha}_t)}\left\|\varepsilon_t^{\mathsf{sc}}(x_t)\right\|_2^2\right\}p_{X_t}(x_t)\mathrm{d}x_t\\
&\quad+\sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}}\eta_t\sup_{\mathcal{A}\subseteq\mathcal{X}_{\mathsf{data}}}\int_{\mathcal{A}}-\Delta\big(\varepsilon_t^{\mathsf{sc}}(x_t),\varepsilon_t^{\mathsf{J}}(x_t)\big)p_{X_t}(x_t)\mathrm{d}x_t.
\end{aligned}
$$

In view of Lemma 6, we know that for any measurable set $\mathcal{A}\subseteq\mathcal{X}_{\mathsf{data}}$,

$$
\begin{aligned}
&\sqrt{\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}}\eta_t\int_{\mathcal{A}} -\Delta\big(\varepsilon_t^{\mathsf{sc}}(x_t),\varepsilon_t^{\mathsf{J}}(x_t)\big)p_{X_t}(x_t)\mathrm{d}x_t\\
&\qquad\leq \frac{2\eta_t}{\sqrt{\alpha_t-\overline{\alpha}_t}}\left\{\varepsilon_{\mathsf{score},t}+\varepsilon_{\mathsf{Jacobi},1,t}+\varepsilon_{\mathsf{Jacobi},2,t}+\varepsilon_{\mathsf{Hess},t}\right\}.
\end{aligned}
$$

Thus, we can further deduce that

$$
\begin{aligned}
\int_{\Phi_t^{-1}(\mathcal{A})\cap\mathcal{E}_t} -W(x_t)p_{X_t}(x_t)\mathrm{d}x_t &\leq \frac{\widetilde{\sigma}_t^2\eta_t}{2(1-\overline{\alpha}_t)}\mathbb{E}_{X_t}\left[\mathrm{tr}(\mathrm{Cov}_{0|t})\right] + \frac{\alpha_t-\overline{\alpha}_t}{1-\overline{\alpha}_t}\widetilde{\sigma}_t^4\mathbb{E}_{X_t}\left[\left\|\mathrm{Cov}_{0|t}\right\|_{\mathrm{F}}^2\right]\\
&\quad+\frac{4(1-\overline{\alpha}_t)\eta_t^2}{\alpha_t-\overline{\alpha}_t}\varepsilon_{\mathsf{Jacobi},1,t}^2 + \frac{\eta_t^2}{2(\alpha_t-\overline{\alpha}_t)}\varepsilon_{\mathsf{score},t}^2\\
&\quad+\frac{2\eta_t}{\sqrt{\alpha_t-\overline{\alpha}_t}}\left\{\varepsilon_{\mathsf{score},t}+\varepsilon_{\mathsf{Jacobi},1,t}+\varepsilon_{\mathsf{Jacobi},2,t}+\varepsilon_{\mathsf{Hess},t}\right\}.
\end{aligned}
\tag{70}
$$

Regarding the second term $\mathbb{P}(X_t \notin \mathcal{E}_t)$, apply Markov's inequality to obtain

$$\begin{aligned}\mathbb{P}(X_t \notin \mathcal{E}_t) &= \mathbb{P}\left(\eta_t \left\|\varepsilon_t^{\mathsf{J}}(X_t)\right\| > \frac{1}{8}\right) \\ &\le 64\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}\eta_t^2\mathbb{E}_{X_t}\left[\left\|\varepsilon_t^{\mathsf{J}}(X_t)\right\|^2\right] \le 64\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}\eta_t^2\mathbb{E}_{X_t}\left[\left\|\varepsilon_t^{\mathsf{J}}(X_t)\right\|_{\mathrm{F}}^2\right] = 64\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}\eta_t^2\varepsilon_{\mathsf{Jacobi},1,t}^2.\end{aligned} \tag{71}$$

Putting (70) and (71) together yields

$$\begin{aligned}&\mathsf{TV}(p_{X_{t-1}}, p_{\Phi_t(X_t)}) = \sup_{\mathcal{A}\subset\mathcal{X}_{\mathsf{data}}}\left\{\mathbb{P}_{X_{t-1}}(\mathcal{A}) - \mathbb{P}_{\Phi_t(X_t)}(\mathcal{A})\right\} \\ &\le \frac{\widetilde{\sigma}_t^2\eta_t}{2(1-\overline{\alpha}_t)}\mathbb{E}_{X_t}\left[\mathrm{tr}(\mathrm{Cov}_{0|t})\right] + \frac{\alpha_t-\overline{\alpha}_t}{1-\overline{\alpha}_t}\widetilde{\sigma}_t^4\mathbb{E}_{X_t}\left[\left\|\mathrm{Cov}_{0|t}\right\|_{\mathrm{F}}^2\right] + \frac{4(1-\overline{\alpha}_t)\eta_t^2}{\alpha_t-\overline{\alpha}_t}\varepsilon_{\mathsf{Jacobi},1,t}^2 + \frac{\eta_t^2}{2(\alpha_t-\overline{\alpha}_t)}\varepsilon_{\mathsf{score},t}^2 \\ &\quad + \frac{2\eta_t}{\sqrt{\alpha_t-\overline{\alpha}_t}}\left\{\varepsilon_{\mathsf{score},t} + \varepsilon_{\mathsf{Jacobi},1,t} + \varepsilon_{\mathsf{Jacobi},2,t} + \varepsilon_{\mathsf{Hess},t}\right\} + 16\frac{1-\overline{\alpha}_t}{\alpha_t-\overline{\alpha}_t}\eta_t^2\varepsilon_{\mathsf{Jacobi},1,t}^2 \\ &\le \frac{\widetilde{\sigma}_t^2\eta_t}{2(1-\overline{\alpha}_t)}\mathbb{E}_{X_t}\left[\mathrm{tr}(\mathrm{Cov}_{0|t})\right] + \frac{\alpha_t-\overline{\alpha}_t}{1-\overline{\alpha}_t}\widetilde{\sigma}_t^4\mathbb{E}_{X_t}\left[\left\|\mathrm{Cov}_{0|t}\right\|_{\mathrm{F}}^2\right] + \frac{20(1-\overline{\alpha}_t)\eta_t^2}{\alpha_t-\overline{\alpha}_t}\varepsilon_{\mathsf{Jacobi},1,t}^2 + \frac{\eta_t^2}{2(\alpha_t-\overline{\alpha}_t)}\varepsilon_{\mathsf{score},t}^2 \\ &\quad + \frac{2\eta_t}{\sqrt{\alpha_t-\overline{\alpha}_t}}\left\{\varepsilon_{\mathsf{score},t} + \varepsilon_{\mathsf{Jacobi},1,t} + \varepsilon_{\mathsf{Jacobi},2,t} + \varepsilon_{\mathsf{Hess},t}\right\}.\end{aligned} \tag{72}$$

Lemma 3 tells us that

$$\frac{\alpha_t-\overline{\alpha}_t}{1-\overline{\alpha}_t}\widetilde{\sigma}_t^4\mathbb{E}\left[\left\|\mathrm{Cov}_{0|t}\right\|_{\mathrm{F}}^2\right] \le \frac{3(\alpha_t-\overline{\alpha}_t)}{1-\overline{\alpha}_t}\widetilde{\sigma}_t^2\left\{\mathbb{E}\left[\mathrm{tr}(\mathrm{Cov}_{X_0|X_t})\right] - \mathbb{E}\left[\mathrm{tr}(\mathrm{Cov}_{X_0|X_{t-1}})\right]\right\} + \frac{1}{T^{10}}. \tag{73}$$

Taking (50), (72) and (73) collectively, we can demonstrate that

$$\begin{aligned}&\mathsf{TV}(p_{X_{t-1}}, p_{Y_{t-1}}) \le \mathsf{TV}(p_{X_{t-1}}, p_{\Phi_t(X_t)}) + \mathsf{TV}(p_{X_t}, p_{Y_t}) \\ &\le \mathsf{TV}(p_{X_t}, p_{Y_t}) + \underbrace{\frac{\widetilde{\sigma}_t^2(\eta_t+3\alpha_t-3\overline{\alpha}_t)}{1-\overline{\alpha}_t}\mathbb{E}\left[\mathrm{tr}\left(\mathrm{Cov}_{0|t}\right)\right] - \frac{3(\alpha_t-\overline{\alpha}_t)}{1-\overline{\alpha}_t}\widetilde{\sigma}_t^2\mathbb{E}\left[\mathrm{tr}\left(\mathrm{Cov}_{0|t-1}\right)\right] + \frac{1}{T^{10}}}_{=:\mathcal{S}_{t,1}} \\ &\quad + \underbrace{\frac{2\eta_t}{\sqrt{\alpha_t-\overline{\alpha}_t}}\left\{\varepsilon_{\mathsf{score},t}+\varepsilon_{\mathsf{Jacobi},1,t}+\varepsilon_{\mathsf{Jacobi},2,t}+\varepsilon_{\mathsf{Hess},t}\right\} + \frac{20(1-\overline{\alpha}_t)\eta_t^2}{\alpha_t-\overline{\alpha}_t}\varepsilon_{\mathsf{Jacobi},1,t}^2 + \frac{\eta_t^2}{2(\alpha_t-\overline{\alpha}_t)}\varepsilon_{\mathsf{score},t}^2}_{=:\mathcal{S}_{t,2}}.\end{aligned} \tag{74}$$

Here, we divide the residual terms generated by recursion into two parts, $\mathcal{S}_{t,1}$ and $\mathcal{S}_{t,2}$. The term $\mathcal{S}_{t,1}$ represents the discretization error, while $\mathcal{S}_{t,2}$ is associated with the score estimation error.

Let us first cope with the accumulated discretization error $\sum_{t=1}^{T}\mathcal{S}_{t,1}$. Simple algebraic transformation yields

$$\begin{aligned}\sum_{t=1}^{T}\mathcal{S}_{t,1} &= \sum_{t=1}^{T-1}\left(\frac{\widetilde{\sigma}_t^2\eta_t}{1-\overline{\alpha}_t} + \frac{3(\alpha_t-\overline{\alpha}_t)\widetilde{\sigma}_t^2}{1-\overline{\alpha}_t} - \frac{3(\alpha_{t+1}-\overline{\alpha}_{t+1})\widetilde{\sigma}_{t+1}^2}{1-\overline{\alpha}_{t+1}}\right)\mathbb{E}\left[\mathrm{tr}\left(\mathrm{Cov}_{0|t}\right)\right] \\ &\quad + \frac{\widetilde{\sigma}_T^2(\eta_T+\alpha_T-\overline{\alpha}_T)}{1-\overline{\alpha}_T}\mathbb{E}\left[\mathrm{tr}\left(\mathrm{Cov}_{0|T}\right)\right] + \frac{1}{T^9}.\end{aligned}$$

Applying Lemma 4 and the basic property (46), we can show that

$$\begin{aligned}\frac{\widetilde{\sigma}_t^2\eta_t}{1-\overline{\alpha}_t} + \frac{3(\alpha_t-\overline{\alpha}_t)\widetilde{\sigma}_t^2}{1-\overline{\alpha}_t} - \frac{3(\alpha_{t+1}-\overline{\alpha}_{t+1})\widetilde{\sigma}_{t+1}^2}{1-\overline{\alpha}_{t+1}} &= \frac{\widetilde{\sigma}_t^2\eta_t}{1-\overline{\alpha}_t} + 3\left(1-\frac{1-\alpha_t}{1-\overline{\alpha}_t}\right)\widetilde{\sigma}_t^2 - 3\left(1-\frac{1-\alpha_{t+1}}{1-\overline{\alpha}_{t+1}}\right)\widetilde{\sigma}_{t+1}^2 \\ &\le \frac{3(1-\alpha_t)\widetilde{\sigma}_t^2}{1-\overline{\alpha}_t} - \frac{3(1-\alpha_t)\widetilde{\sigma}_t^2}{1-\overline{\alpha}_t} + \frac{3(1-\alpha_{t+1})\widetilde{\sigma}_{t+1}^2}{1-\overline{\alpha}_{t+1}} + 3(\widetilde{\sigma}_t^2 - \widetilde{\sigma}_{t+1}^2)\end{aligned}$$

$$\leq 3\left(\frac{1-\alpha_{t+1}}{1-\overline{\alpha}_{t+1}}\right)^2 \frac{\overline{\alpha}_{t+1}^2}{1-\overline{\alpha}_{t+1}} + \frac{C_6 \log^2 T}{T^2}\frac{\overline{\alpha}_t}{1-\overline{\alpha}_t} \leq \frac{2C_6 \log^2 T}{T^2}\frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}.$$

Equipped with this, the accumulated discretization error can be further bounded as follows:

$$\begin{aligned}\sum_{t=1}^{T}\mathcal{S}_{t,1} &\leq \left(\frac{2C_6\log T}{T}\right)^2 \sum_{t=1}^{T-1}\frac{\overline{\alpha}_t}{1-\overline{\alpha}_t}\mathbb{E}\left[\mathrm{tr}\left(\mathrm{Cov}_{0|t}\right)\right] + \frac{2C_6\log T}{T}\frac{\overline{\alpha}_T}{1-\overline{\alpha}_T}\mathbb{E}\left[\mathrm{tr}\left(\mathrm{Cov}_{0|T}\right)\right] + \frac{1}{T^9} \\ &\overset{(a)}{\leq} C_3 kT\log T\left(\frac{2C_6\log T}{T}\right)^2 + \frac{2C_3C_6 k\log^2 T}{T} + \frac{1}{T^9} \leq C_9\frac{k\log^3 T}{T} + \frac{1}{T^9} \leq C_{10}\frac{k\log^3 T}{T},\end{aligned} \tag{75}$$

where (a) applies the moment inequality (45) with $l = 2$.

Next, we turn to the cumulative estimation error $\sum_{t=1}^{T}\mathcal{S}_{t,2}$. Given that $\frac{\eta_t}{\sqrt{\alpha_t-\overline{\alpha}_t}} \leq \frac{\eta_t}{1-\overline{\alpha}_t} \leq \frac{1-\alpha_t}{1-\overline{\alpha}_t} \leq \frac{8c_1\log T}{T}$, we can derive

$$\begin{aligned}\sum_{t=2}^{T}\mathcal{S}_{t,2} &= \sum_{t=2}^{T}\frac{2\eta_t}{\sqrt{\alpha_t-\overline{\alpha}_t}}\left\{\varepsilon_{\mathsf{score},t}+\varepsilon_{\mathsf{Jacobi},1,t}+\varepsilon_{\mathsf{Jacobi},2,t}+\varepsilon_{\mathsf{Hess},t}\right\} \\ &\quad + \sum_{t=2}^{T}\frac{20(1-\overline{\alpha}_t)\eta_t^2}{\alpha_t-\overline{\alpha}_t}\varepsilon_{\mathsf{Jacobi},1,t}^2 + \sum_{t=2}^{T}\frac{\eta_t^2}{2(\alpha_t-\overline{\alpha}_t)}\varepsilon_{\mathsf{score},t}^2 \\ &\leq \frac{8c_1\log T}{T}\left\{\sum_{t=1}^{T}\varepsilon_{\mathsf{score},t}+\sum_{t=2}^{T}\varepsilon_{\mathsf{Jacobi},1,t}+\sum_{t=2}^{T}\varepsilon_{\mathsf{Jacobi},2,t}+\sum_{t=2}^{T}\varepsilon_{\mathsf{Hess},t}\right\} \\ &\quad + \frac{C_{10}\log^2 T}{T^2}\left\{\sum_{t=1}^{T}\varepsilon_{\mathsf{Jacobi},1,t}^2+\sum_{t=1}^{T}\varepsilon_{\mathsf{score},t}^2\right\} \\ &\overset{(a)}{\leq} \frac{8c_1\log T}{\sqrt{T}}\left\{\sqrt{\sum_{t=2}^{T}\varepsilon_{\mathsf{score},t}^2}+\sqrt{\sum_{t=2}^{T}\varepsilon_{\mathsf{Jacobi},1,t}^2}+\sqrt{\sum_{t=2}^{T}\varepsilon_{\mathsf{Jacobi},2,t}^2}+\sqrt{\sum_{t=2}^{T}\varepsilon_{\mathsf{Hess},t}^2}\right\} \\ &\quad + \frac{C_{10}\log^2 T}{T^2}\left\{\sum_{t=2}^{T}\varepsilon_{\mathsf{Jacobi},1,t}^2+\sum_{t=2}^{T}\varepsilon_{\mathsf{score},t}^2\right\} \\ &\overset{(b)}{\leq} 8c_1\left(\varepsilon_{\mathsf{score}}+\varepsilon_{\mathsf{Jacobi},1}+\varepsilon_{\mathsf{Jacobi},2}+\varepsilon_{\mathsf{Hess}}\right)\log T + \frac{C_{10}\log^2 T}{T}\left(\varepsilon_{\mathsf{score}}^2+\varepsilon_{\mathsf{Jacobi},1}^2\right) \\ &\leq C_{11}\left(\varepsilon_{\mathsf{score}}+\varepsilon_{\mathsf{Jacobi},1}+\varepsilon_{\mathsf{Jacobi},2}+\varepsilon_{\mathsf{Hess}}\right)\log T,\end{aligned} \tag{76}$$

where (a) results from the Cauchy-Schwarz inequality, (b) follows from Assumption 4, and the last inequality holds provided that $\frac{\log T}{T}(\varepsilon_{\mathsf{score}}+\varepsilon_{\mathsf{Jacobi},1}) \leq 1$.

Applying inequality (74) recursively from 1 to T, and combining (75) and (76), we reach

$$\begin{aligned}\mathsf{TV}(p_{X_1}, p_{Y_1}) &\leq \mathsf{TV}(p_{X_T}, p_{Y_T}) + \sum_{t=2}^{T}\mathcal{S}_{t,1} + \sum_{t=2}^{T}\mathcal{S}_{t,2} \\ &\leq C_{10}\frac{k\log^3 T}{T} + C_{11}\left(\varepsilon_{\mathsf{score}}+\varepsilon_{\mathsf{Jacobi},1}+\varepsilon_{\mathsf{Jacobi},2}+\varepsilon_{\mathsf{Hess}}\right)\log T + \mathsf{TV}(p_{X_T}, p_{Y_T}) \\ &\leq C_{10}\frac{k\log^3 T}{T} + C_{11}\left(\varepsilon_{\mathsf{score}}+\varepsilon_{\mathsf{Jacobi},1}+\varepsilon_{\mathsf{Jacobi},2}+\varepsilon_{\mathsf{Hess}}\right)\log T + \frac{1}{T^{10}},\end{aligned}$$

where the last inequality arises from Li and Yan (2024a, Lemma 10). The proof is thus complete.

C Analysis for DDPM (proof of Theorem 3)

Given that Theorem 2 is a special case of Theorem 3, we shall focus on proving Theorem 3 in this section.

Define the extended d-dimensional Euclidean space $\mathbb{R}^d \cup \{\infty\}$ by adding the point ∞ to $\mathbb{R}^d$. From now on, the random vectors can take value in $\mathbb{R}^d \cup \{\infty\}$, namely, they can be constructed in the following way:

$$X = \begin{cases} X', & \text{with probability } \theta, \\ \infty, & \text{with probability } 1-\theta, \end{cases}$$

where $\theta \in [0,1]$ and X' is a random vector in $\mathbb{R}^d$ in the usual sense. If X' has a density $p_{X'}$, then the generalized density of X is

$$p_X(x) = \theta p_{X'}(x)\mathbb{1}\{x \in \mathbb{R}^d\} + (1-\theta)\delta_\infty,$$

where δ_∞ indicates the Dirac measure at ∞.

C.1 Introducing auxiliary sequences

Let us define two auxiliary sequences that play a pivotal role in our theoretical analysis. First, define an auxiliary reverse process $\overline{Y}_t$ based on Y_t^* as follows:

$$\overline{Y}_{t-1}^- \mid \overline{Y}_t \sim Y_{t-1}^* \mid Y_t^*, \qquad \overline{Y}_t \mid \{\overline{Y}_t^- = y_t\} = \begin{cases} y_t^-, & \text{with prob. } \frac{p_{X_t}(y_t^-)}{p_{\overline{Y}_t^-}(y_t^-)} \wedge 1, \\ \infty, & \text{otherwise.} \end{cases}$$

Here, $a \wedge b \coloneqq \min\{a, b\}$. One can derive the following property of the auxiliary sequence $\{\overline{Y}_t\}_{t=1}^\infty$:

$$p_{\overline{Y}_t}(y_t) = \int_{\mathbb{R}^d} p_{\overline{Y}_t|\overline{Y}_t^-}(y_t \mid y_t^-)\mathrm{d}y_t^- = p_{X_t}(y_t) \wedge p_{\overline{Y}_t^-}(y_t). \tag{77}$$

To control the estimation error, we introduce another auxiliary reverse process $\widehat{Y}_t$ based on Y_t:

$$\widehat{Y}_{t-1}^- \mid \widehat{Y}_t \sim Y_{t-1} \mid Y_t, \qquad \widehat{Y}_t \mid \{\widehat{Y}_t^- = y_t\} = \begin{cases} y_t^-, & \text{with prob. } \frac{p_{X_t}(y_t^-)}{p_{\overline{Y}_t^-}(y_t^-)} \wedge 1, \\ \infty, & \text{otherwise.} \end{cases}$$

We can use induction to show that

$$p_{Y_t}(x) \geq p_{\widehat{Y}_t}(x), \qquad \forall x \in \mathcal{X}_{\mathsf{data}} \tag{78}$$

holds for all $t = 1, \cdots, T$. In fact, we can check that $p_{Y_T} = p_{\widehat{Y}_T}$. Suppose (78) holds for $t+1$, then

$$\begin{aligned} p_{\widehat{Y}_t}(x) &= \int_{\mathbb{R}^d} p_{\widehat{Y}_t|\widehat{Y}_t^-}(x \mid x')p_{\widehat{Y}_t^-}(x')\mathrm{d}x' = \left(\frac{p_{X_t}(x)}{p_{\overline{Y}_t^-}(x)} \wedge 1\right) p_{\widehat{Y}_t^-}(x) \leq p_{\widehat{Y}_t^-}(x) \\ &= \int_{\mathbb{R}^d} p_{\widehat{Y}_t^-|\widehat{Y}_{t+1}}(x \mid x')p_{\widehat{Y}_{t+1}}(x')\mathrm{d}x' \leq \int_{\mathbb{R}^d} p_{Y_t|Y_{t+1}}(x \mid x')p_{Y_{t+1}}(x')\mathrm{d}x' = p_{Y_t}(x). \end{aligned}$$

In order to facilitate discussion, we define the following set

$$\mathcal{A}_t \coloneqq \left\{x : p_{X_t}(x) \geq p_{\overline{Y}_t^-}(x)\right\}.$$

It can be easily verified that

$$\mathsf{TV}(p_{X_t}, p_{Y_t^*}) = \int_{\mathcal{A}_t} \big(p_{X_t}(x_t) - p_{Y_t^*}(x)\big)\mathrm{d}x.$$

Note that, for density $p_{\overline{Y}_{t-1}^-}(x_{t-1})$, we have

$$p_{\overline{Y}_{t-1}}(x_{t-1}) = \int_{\mathcal{X}_{\mathsf{data}}} p_{\overline{Y}_{t-1}^-|\overline{Y}_t}(x_{t-1} \mid x_t)p_{\overline{Y}_t}(x_t)\mathrm{d}x_t = \int_{\mathcal{X}_{\mathsf{data}}} p_{Y_{t-1}^*|Y_t^*}(x_{t-1} \mid x_t)p_{\overline{Y}_t}(x_t)\mathrm{d}x_t$$

$$= \int_{\mathcal{X}_{\mathsf{data}}} p_{Y^*_{t-1}|Y^*_t}(x_{t-1} \mid x_t) p_{X_t}(x_t) \mathrm{d}x_t + \int_{\mathcal{X}_{\mathsf{data}}} p_{Y^*_{t-1}|Y^*_t}(x_{t-1} \mid x_t)(p_{\overline{Y}_t}(x_t) - p_{X_t}(x_t)) \mathrm{d}x_t.$$

Further, by plugging the above equation into the expression of the TV distance between p_{X_t} and $p_{Y^*_t}$, we can derive the following recursive relationship for all $t \geq 2$:

$$\begin{aligned}
\mathsf{TV}(p_{X_{t-1}}, p_{\overline{Y}_{t-1}}) &= \int_{\mathcal{A}_{t-1}} \big(p_{X_{t-1}}(x_{t-1}) - p_{\overline{Y}_{t-1}}(x_{t-1})\big) \mathrm{d}x_{t-1} \overset{(a)}{=} \int_{\mathcal{A}_{t-1}} \big(p_{X_{t-1}}(x_{t-1}) - p_{\overline{Y}^-_{t-1}}(x_{t-1})\big) \mathrm{d}x_{t-1} \\
&= \underbrace{\int_{\mathcal{A}_{t-1}} p_{X_{t-1}}(x_{t-1}) - \int_{\mathcal{A}_{t-1} \times \mathcal{X}_{\mathsf{data}}} p_{Y^*_{t-1}|Y^*_t}(x_{t-1} \mid x_t) p_{X_t}(x_t) \mathrm{d}x_{t-1} \mathrm{d}x_t}_{=: \mathcal{R}_{t-1}} \\
&\quad + \int_{\mathcal{A}_{t-1} \times \mathcal{X}_{\mathsf{data}}} p_{Y^*_{t-1}|Y^*_t}(x_{t-1} \mid x_t) \big(p_{X_t}(x_t) - p_{\overline{Y}_t}(x_t)\big) \mathrm{d}x_{t-1} \mathrm{d}x_t \\
&\overset{(b)}{\leq} \mathcal{R}_{t-1} + \int_{\mathcal{A}_{t-1} \times \mathcal{A}_t} p_{Y^*_{t-1}|Y^*_t}(x_{t-1} \mid x_t) \big(p_{X_t}(x_t) - p_{\overline{Y}_t}(x_t)\big) \mathrm{d}x_{t-1} \mathrm{d}x_t \\
&= \mathcal{R}_{t-1} + \int_{\mathcal{A}_t} \mathbb{P}_{Y^*_{t-1}|Y^*_t}(\mathcal{A}_{t-1}|x_t) \big(p_{X_t}(x_t) - p_{\widehat{Y}_t}(x_t)\big) \mathrm{d}x_t \overset{(c)}{\leq} \mathcal{R}_{t-1} + \mathsf{TV}(p_{X_t}, p_{\overline{Y}_t}).
\end{aligned} \tag{79}$$

Here, (a) follows since $p_{\overline{Y}_{t-1}} = p_{\overline{Y}^-_{t-1}}$ on $\mathcal{A}_{t-1}$, (b) holds due to the fact that $p_{Y^*_{t-1}|Y^*_t}(x_{t-1} \mid x_t)(p_{X_t}(x_t) - p_{Y^*_t}(x_t)) < 0$ on $\mathcal{A}^c_t$, and (c) is valid since $\mathbb{P}_{Y^*_{t-1}|Y^*_t}(\mathcal{A}_{t-1} \mid x_t) \leq 1$ for all $x_t \in \mathcal{X}_{\mathsf{data}}$. Denote $p_{X_{t-1}}(x_{t-1}) - \int_{\mathcal{X}} p_{Y^*_{t-1}|Y^*_t}(x_{t-1} \mid x_t) p_{X_t}(x_t) \mathrm{d}x_t$ as $\mathcal{R}_{t-1}(x_{t-1})$, then we can see that

$$\mathcal{R}_{t-1} = \int_{\mathcal{A}_{t-1}} \mathcal{R}_{t-1}(x_{t-1}) \mathrm{d}x_{t-1}.$$

In the ensuing subsections, we would like to bound $\mathcal{R}_{t-1}(x_{t-1})$ for $x_{t-1} \in \mathcal{A}_{t-1}$, which can then be invoked to bound $\mathcal{R}_{t-1}$.

C.2 Analysis of $\mathcal{R}_{t-1}(x_{t-1})$

Recall that $Y^*_{t-1} = \frac{1}{\sqrt{\alpha_t}}\big(Y^*_t + \eta_t s^*_t(Y^*_t) + \sigma_t W_t\big)$. For the vector x_t, denote

$$u_t \coloneqq x_t + \eta_t s^*_t(x_t).$$

Note that for $t \geq 2$,

$$\begin{aligned}
p_{X_{t-1}|X_0}(x_{t-1} \mid x_0) &= \left(\frac{1}{2\pi(1-\overline{\alpha}_t)}\right)^{\frac{d}{2}} \exp\left(-\frac{\|x_{t-1} - \sqrt{\overline{\alpha}_{t-1}} x_0\|_2^2}{2(1-\overline{\alpha}_{t-1})}\right), \\
p_{Y^*_{t-1}|Y^*_t}(x_{t-1} \mid u_t) &= \left(\frac{1}{2\pi\sigma_t^2}\right)^{\frac{d}{2}} \exp\left(-\frac{\left\|\sqrt{\alpha_t} x_{t-1} - u_t\right\|_2^2}{2\sigma_t^2}\right),
\end{aligned}$$

which means both $p_{X_{t-1}|X_0}$ and $p_{Y^*_{t-1}|Y^*_t}$ are density functions of conditional Gaussians. In light of this, we can find another conditional Gaussian distribution $\widetilde{p}_{U_t|X_0}(u_t|x_0)$ such that the following convolutional formula holds

$$p_{X_{t-1}|X_0}(x_{t-1} \mid x_0) = \int_{x_t} p_{Y^*_{t-1}|Y^*_t}(x_{t-1} \mid u_t) \widetilde{p}_{U_t|X_0}(u_t \mid x_0) \mathrm{d}u_t, \quad \forall x_{t-1} \in \mathcal{A}_{t-1},\ x_0 \in \mathcal{X}_{\mathsf{data}}. \tag{80}$$

Further, we can derive

$$\mathcal{R}_{t-1}(x_{t-1}) = p_{X_{t-1}}(x_{t-1}) - \int_{x_0, x_t} p_{Y^*_{t-1}|Y^*_t}(x_{t-1} \mid x_t) p_{X_t|X_0}(x_t \mid x_0) p(x_0) \mathrm{d}x_0 \mathrm{d}x_t$$

$$
\begin{aligned}
&= p_{X^*_{t-1}}(x_{t-1}) - \int_{x_0,x_t} p_{Y^*_{t-1}|Y^*_t}(x_{t-1} \mid u_t)\widetilde{p}_{U_t|X_0}(u_t \mid x_0)\frac{p_{X_t|X_0}(u_t \mid x_0)}{\widetilde{p}_{U_t|X_0}(u_t \mid x_0)}p(x_0)\mathrm{d}x_0\mathrm{d}x_t \\
&= p_{X^*_{t-1}}(x_{t-1}) - \int_{x_0,x_t} p_{Y^*_{t-1}|Y^*_t}(x_{t-1} \mid u_t)\widetilde{p}_{U_t|X_0}(u_t \mid x_0)\frac{p_{X_t|X_0}(u_t \mid x_0)}{\widetilde{p}_{U_t|X_0}(u_t \mid x_0)}\det\left(\frac{\mathrm{d}x_t}{\mathrm{d}u_t}\right)p(x_0)\mathrm{d}x_0\mathrm{d}u_t \\
&= \left(\int_{x_0} p_{X_{t-1}|X_0}(x_{t-1} \mid x_0)p(x_0)\mathrm{d}x_0 - \int_{x_0,x_t} p_{Y^*_{t-1}|Y^*_t}(x_{t-1} \mid u_t)\widetilde{p}_{U_t|X_0}(u_t \mid x_0)p(x_0)\mathrm{d}x_0\mathrm{d}u_t\right) \\
&\quad - \int_{x_0,x_t} p_{Y^*_{t-1}|Y^*_t}(x_{t-1} \mid u_t)\widetilde{p}_{U_t|X_0}(u_t \mid x_0)\left(\frac{p_{X_t|X_0}(u_t \mid x_0)}{\widetilde{p}_{U_t|X_0}(u_t \mid x_0)}\det\left(\frac{\mathrm{d}x_t}{\mathrm{d}u_t}\right) - 1\right)p(x_0)\mathrm{d}x_0\mathrm{d}u_t.
\end{aligned}
$$

According to (80), the value in parentheses at the end of the above equation is 0. In addition, combining the definition of u_t leads to

$$
\begin{aligned}
&- \int_{x_0,x_t} p_{Y^*_{t-1}|Y^*_t}(x_{t-1} \mid u_t)\widetilde{p}_{U_t|X_0}(u_t \mid x_0)\left(\frac{p_{X_t|X_0}(u_t \mid x_0)}{\widetilde{p}_{U_t|X_0}(u_t \mid x_0)}\det\left(\frac{\mathrm{d}x_t}{\mathrm{d}u_t}\right) - 1\right)p(x_0)\mathrm{d}x_0\mathrm{d}u_t \\
&= - \int_{x_0,x_t} p_{Y^*_{t-1}|Y^*_t}(x_{t-1} \mid u_t)p_{X_t|X_0}(u_t \mid x_0)\det\left(\frac{\mathrm{d}x_t}{\mathrm{d}u_t}\right)\left(1 - \frac{\widetilde{p}_{U_t|X_0}(u_t \mid x_0)}{p_{X_t|X_0}(u_t \mid x_0)}\det\left(\frac{\mathrm{d}u_t}{\mathrm{d}x_t}\right)\right)p(x_0)\mathrm{d}x_0\mathrm{d}u_t \\
&= \int_{x_0,x_t} p_{Y^*_{t-1}|Y^*_t}(x_{t-1} \mid u_t)p_{X_t|X_0}(u_t \mid x_0)\left(\frac{\widetilde{p}_{U_t|X_0}(u_t \mid x_0)}{p_{X_t|X_0}(u_t \mid x_0)}\det\left(\frac{\mathrm{d}u_t}{\mathrm{d}x_t}\right) - 1\right)p(x_0)\mathrm{d}x_0\mathrm{d}x_t.
\end{aligned}
$$

Let

$$
\mathcal{G}(x_t \mid x_0) \coloneqq \frac{\widetilde{p}_{U_t|X_0}(u_t \mid x_0)}{p_{X_t|X_0}(u_t \mid x_0)}\det\left(\frac{\mathrm{d}u_t}{\mathrm{d}x_t}\right),
$$

then we can derive the following equation via the above analysis

$$
\mathcal{R}_{t-1}(x_{t-1}) = \int_{x_0,x_t} p_{Y^*_{t-1}|Y^*_t}(x_{t-1} \mid u_t)p_{X_t|X_0}(u_t \mid x_0)\big(\mathcal{G}(x_t,x_0) - 1\big)p(x_0)\mathrm{d}x_0\mathrm{d}x_t. \tag{81}
$$

Next, we have two steps to accomplish. The first one is to determine the specific form of the surrogate distribution $\widetilde{p}_{U_t|X_0}(u_t \mid x_0)$, while the second one is to control the magnitude of $\mathcal{G}(x_t,x_0) - 1$. These are detailed in the subsequent subsections.

C.3 Determining the surrogate distribution

Suppose that $\widetilde{p}_{U_t|X_0}(u_t|x_0)$ satisfies $U_t \sim \mathcal{N}(\lambda_t x_0, \overline{\sigma}_t^2 I)$. It is easily seen that the distribution of the probabilistic density function $\int_{x_t} p_{Y^*_{t-1}|Y^*_t}(x_{t-1}|u_t)\widetilde{p}_{U_t|X_0}(u_t|x_0)\mathrm{d}u_t$ is

$$
X_{t-1} \sim \mathcal{N}\left(\frac{\lambda_t}{\sqrt{\alpha_t}}x_0,\ \frac{\sigma_t^2 + \overline{\sigma}_t^2}{\alpha_t}\right),
$$

while under the choice of $\widetilde{p}_{U_t|X_0}$ we have

$$
p_{X_{t-1}|X_0}(x_{t-1} \mid x_0) = \int_{x_t} p_{Y^*_{t-1}|Y^*_t}(x_{t-1} \mid u_t)\widetilde{p}_{U_t|X_0}(u_t \mid x_0)\mathrm{d}u_t, \qquad \forall x_{t-1} \in \mathcal{A}_{t-1},\ x_0 \in \mathcal{X}_{\mathsf{data}}.
$$

This implies that $\mathcal{N}\left(\frac{\lambda_t}{\sqrt{\alpha_t}}x_0,\ \frac{\sigma_t^2+\overline{\sigma}_t^2}{\alpha_t}\right)$ is the same as $\mathcal{N}\left(\sqrt{\overline{\alpha}_{t-1}}x_0, 1-\overline{\alpha}_{t-1}\right)$. Combining this with the choice of σ_t^2 leads to

$$
\lambda_t = \sqrt{\alpha_t}\cdot\sqrt{\overline{\alpha}_{t-1}} = \sqrt{\overline{\alpha}_t}; \quad \overline{\sigma}_t^2 = \alpha_t(1-\overline{\alpha}_{t-1}) - \sigma_t^2.
$$

Hence, the surrogate distribution of density $\widetilde{p}_{U_t|X_0}$ is $\mathcal{N}\left(\sqrt{\overline{\alpha}_t}x_0, \alpha_t - \overline{\alpha}_t - \sigma_t^2\right)$.

C.4 Bounding the residual $\mathcal{G}(x_t, x_0) - 1$

Since both $\widetilde{p}_{U_t|X_0}$ and $p_{X_t|X_0}$ are Gaussian distributions, we can derive that

$$\begin{aligned}\mathcal{G}(x_t, x_0) &= \det\left(\frac{\mathrm{d}(x_t + \eta_t s_t^*(x_t))}{\mathrm{d}x_t}\right) \cdot \frac{\left(\overline{\sigma}_t^2\right)^{-d/2} \exp\left\{-\frac{\left\|u_t - \sqrt{\overline{\alpha}_t} x_0\right\|_2^2}{2\overline{\sigma}_t^2}\right\}}{(1-\overline{\alpha}_t)^{-d/2} \exp\left\{-\frac{\left\|x_t - \sqrt{\overline{\alpha}_t} x_0\right\|^2}{2(1-\overline{\alpha}_t)}\right\}} \\ &= \underbrace{\det\left(\frac{\mathrm{d}(x_t + \eta_t s_t^*(x_t))}{\mathrm{d}x_t}\right) \cdot \frac{(1-\overline{\alpha}_t)^{d/2}}{\overline{\sigma}_t^d}}_{=:\,\mathcal{G}_1(x_t, x_0)} \cdot \underbrace{\exp\left\{\frac{\left\|x_t - \sqrt{\overline{\alpha}_t} x_0\right\|_2^2}{2(1-\overline{\alpha}_t)} - \frac{\left\|u_t - \sqrt{\overline{\alpha}_t} x_0\right\|_2^2}{2\overline{\sigma}_t^2}\right\}}_{=:\,\mathcal{G}_2(x_t, x_0)}.\end{aligned} \tag{82}$$

From Tweedie's formula (42), we can further simplify $\mathcal{G}_1(x_t, x_0)$ as follows:

$$\begin{aligned}\mathcal{G}_1(x_t, x_0) &= \det\left(I + \eta_t \frac{\partial s_t^*(x_t)}{\partial x_t}\right) \cdot \frac{(1-\overline{\alpha}_t)^{d/2}}{\overline{\sigma}_t^d} \\ &\overset{(a)}{=} \det\left(I + \eta_t \left\{\frac{\overline{\alpha}_t}{(1-\overline{\alpha}_t)^2} \mathrm{Cov}_{0|t}(x_t) - \frac{1}{1-\overline{\alpha}_t} I\right\}\right) \cdot \frac{(1-\overline{\alpha}_t)^{d/2}}{\overline{\sigma}_t^d} \\ &= \det\left(\left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right) I + \frac{\overline{\alpha}_t \eta_t}{(1-\overline{\alpha}_t)^2} \mathrm{Cov}_{0|t}(x_t)\right) \cdot \frac{(1-\overline{\alpha}_t)^{d/2}}{\overline{\sigma}_t^d} \\ &\overset{(b)}{=} \det\left(I + \frac{\overline{\alpha}_t \eta_t}{(1-\overline{\alpha}_t)(1-\overline{\alpha}_t-\eta_t)} \mathrm{Cov}_{0|t}(x_t)\right) =: \det\left(I + \overline{\alpha}_t \eta_t^{(1)} \mathrm{Cov}_{0|t}(x_t)\right).\end{aligned} \tag{83}$$

Here, $\eta_t^{(1)} := \frac{\eta_t}{(1-\overline{\alpha}_t)(1-\overline{\alpha}_t-\eta_t)}$. Eqn. (a) follows from the expression of $\frac{\partial s_t^*(x_t)}{\partial x_t}$ listed above, whereas (b) holds by combining $\det(\lambda A) = \lambda^d \det(A)$ and the relationship (26).

We then move on to $\mathcal{G}_2(x_t, x_0)$. Towards this end, we make the observation that

$$\begin{aligned}\log \mathcal{G}_2(x_t, x_0) &= \frac{\left\|x_t - \sqrt{\overline{\alpha}_t} x_0\right\|_2^2}{2(1-\overline{\alpha}_t)} - \frac{\left\|u_t - \sqrt{\overline{\alpha}_t} x_0\right\|_2^2}{2\overline{\sigma}_t^2} = \frac{\left\|x_t - \sqrt{\overline{\alpha}_t} x_0\right\|_2^2}{2(1-\overline{\alpha}_t)} - \frac{\left\|x_t - \sqrt{\overline{\alpha}_t} x_0 + \eta_t s_t^*(x_t)\right\|_2^2}{2\overline{\sigma}_t^2} \\ &= \left(\frac{1}{2(1-\overline{\alpha}_t)} - \frac{1}{2\overline{\sigma}_t^2}\right) \left\|x_t - \sqrt{\overline{\alpha}_t} x_0\right\|_2^2 - \frac{\eta_t}{\overline{\sigma}_t^2} (x_t - \sqrt{\overline{\alpha}_t} x_0)^\top s_t^*(x_t) - \frac{\eta_t^2}{2\overline{\sigma}_t^2} \left\|s_t^*(x_t)\right\|_2^2 \\ &\overset{(a)}{=} -\frac{\eta_t}{\overline{\sigma}_t^2 (1-\overline{\alpha}_t)} (x_t - \sqrt{\overline{\alpha}_t} x_0)^\top \left(\sqrt{\overline{\alpha}_t} \mu_{0|t}(x_t) - x_t\right) - \frac{\eta_t^2}{2\overline{\sigma}_t^2 (1-\overline{\alpha}_t)^2} \left\|\sqrt{\overline{\alpha}_t} \mu_{0|t}(x_t) - x_t\right\|_2^2 \\ &\quad + \left(\frac{1}{2(1-\overline{\alpha}_t)} - \frac{1}{2\overline{\sigma}_t^2}\right) \left\|x_t - \sqrt{\overline{\alpha}_t} x_0\right\|_2^2 \\ &= \left(\frac{1}{2(1-\overline{\alpha}_t)} - \frac{1}{2\overline{\sigma}_t^2}\right) \left\|x_t - \sqrt{\overline{\alpha}_t} x_0\right\|_2^2 + \left(\frac{\eta_t}{\overline{\sigma}_t^2 (1-\overline{\alpha}_t)} - \frac{\eta_t^2}{2\overline{\sigma}_t^2 (1-\overline{\alpha}_t)^2}\right) \left\|\sqrt{\overline{\alpha}_t} \mu_{0|t}(x_t) - x_t\right\|_2^2 \\ &\quad + \frac{\eta_t}{\overline{\sigma}_t^2 (1-\overline{\alpha}_t)} \left(\sqrt{\overline{\alpha}_t} x_0 - \sqrt{\overline{\alpha}_t} \mu_{0|t}(x_t)\right)^\top \left(\sqrt{\overline{\alpha}_t} \mu_{0|t}(x_t) - x_t\right),\end{aligned} \tag{84}$$

where (a) follows since $s_t^*(x_t) = \frac{\sqrt{\overline{\alpha}_t}}{1-\overline{\alpha}_t} \mu_{0|t}(x_t) - \frac{1}{1-\overline{\alpha}_t} x_t$. Further, in view of (26), one has

$$\begin{aligned}\frac{1}{2(1-\overline{\alpha}_t)} - \frac{1}{2\overline{\sigma}_t^2} &= \frac{\overline{\sigma}_t^2 - (1-\overline{\alpha}_t)}{2(1-\overline{\alpha}_t)\overline{\sigma}_t^2} = \frac{(1-\overline{\alpha}_t)\left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right)^2 - (1-\overline{\alpha}_t)}{2(1-\overline{\alpha}_t)\overline{\sigma}_t^2} \\ &= -\frac{2(1-\overline{\alpha}_t)\eta_t - \eta_t^2}{2(1-\overline{\alpha}_t)^2 \overline{\sigma}_t^2} = -\left(\frac{\eta_t}{(1-\overline{\alpha}_t)\overline{\sigma}_t^2} - \frac{\eta_t^2}{2(1-\overline{\alpha}_t)^2 \overline{\sigma}_t^2}\right).\end{aligned}$$

Denote $\frac{\eta_t}{(1-\overline{\alpha}_t)\overline{\sigma}_t^2} - \frac{\eta_t^2}{2(1-\overline{\alpha}_t)^2\overline{\sigma}_t^2}$ as $\eta_t^{(2)}$. Substituting these equations into (84) yields

$$\begin{aligned}\log \mathcal{G}_2(x_t, x_0) &= \eta_t^{(2)} \left\{ \left\| \sqrt{\overline{\alpha}_t}\mu_{0|t}(x_t) - x_t \right\|_2^2 - \left\| \sqrt{\overline{\alpha}_t} x_0 - x_t \right\|_2^2 \right\} + \frac{\sqrt{\overline{\alpha}_t}\eta_t}{\overline{\sigma}_t^2(\alpha_t - \overline{\alpha}_t)^2} \left(x_0 - \mu_{0|t}(x_t)\right)^\top \left(\sqrt{\overline{\alpha}_t}\mu_{0|t}(x_t) - x_t\right) \\ &=: \zeta(x_t, x_0) - \int_{x_0} \log \mathcal{G}_2(x_t, x_0) p_{X_0|X_t}(x_0 \mid x_t) \mathrm{d}x_0 = \zeta(x_t, x_0) - \overline{\alpha}_t \eta_t^{(2)} \mathrm{tr}\left(\mathrm{Cov}_{0|t}(x_t)\right).\end{aligned} \tag{85}$$

Here, $\zeta(x_t, x_0)$ satisfies $\int_{x_0} \zeta(x_t, x_0) p_{X_0|X_t}(x_0 \mid x_t)\mathrm{d}x_0 = 0,\ \forall x_t \in \mathcal{X}_{\mathsf{data}}$. Now, to streamline the presentation of the proof, we define

$$Z(x_t) := \log\det\left(I + \overline{\alpha}_t \eta_t^{(1)} \mathrm{Cov}_{0|t}(x_t)\right) - \overline{\alpha}_t \eta_t^{(2)} \mathrm{tr}\left(\mathrm{Cov}_{0|t}(x_t)\right).$$

It is readily seen that, for any $x_t \in \mathcal{X}$,

$$\int_{x_0} \left(e^{\zeta(x_t, x_0)} - 1\right) e^{Z(x_t)} p_{X_0|X_t}(x_0 \mid x_t)\mathrm{d}x_0 \geq e^{Z(x_t)} \int_{x_0} \zeta(x_t, x_0) p_{X_0|X_t}(x_0 \mid x_t)\mathrm{d}x_0 = 0. \tag{86}$$

We now proceed to the analysis of $\mathcal{R}(x_{t-1})$. With (81) in mind, we can obtain

$$\begin{aligned}\mathcal{R}_{t-1} &= \int_{\mathcal{A}_{t-1}} \mathcal{R}_{t-1}(x_{t-1})\mathrm{d}x_{t-1} = \int_{\mathcal{A}_{t-1}\times\mathcal{X}_{\mathsf{data}}^2} p_{Y_{t-1}^*|Y_t^*}(x_{t-1}|u_t) p_{X_t,X_0}(x_t, x_0) \left(\mathcal{G}(x_t, x_0) - 1\right) \mathrm{d}x_0 \mathrm{d}x_t \mathrm{d}x_{t-1} \\ &= \int_{\mathcal{A}_{t-1}\times\mathcal{X}_{\mathsf{data}}^2} p_{Y_{t-1}^*|Y_t^*}(x_{t-1}|u_t) p_{X_t,X_0}(x_t, x_0) \left\{\mathcal{G}_1(x_t, x_0)\mathcal{G}_2(x_t, x_0) - 1\right\} \mathrm{d}x_0 \mathrm{d}x_t \mathrm{d}x_{t-1} \\ &\overset{(a)}{=} \int_{\mathcal{A}_{t-1}\times\mathcal{X}_{\mathsf{data}}^2} p_{Y_{t-1}^*|Y_t^*}(x_{t-1}|u_t) p_{X_t,X_0}(x_t, x_0) \left\{e^{\zeta(x_t, x_0) + Z(x_t)} - 1\right\} \mathrm{d}x_0 \mathrm{d}x_t \mathrm{d}x_{t-1} \\ &\leq \int_{x_t} \mathbb{P}_{Y_{t-1}^*|Y_t^*}\left(\mathcal{A}_{t-1}|u_t\right) p_{X_t}(x_t) \left\{\int_{x_0} (e^{\zeta(x_t, x_0)} - 1) e^{Z(x_t)} p_{X_0|X_t}(x_0 \mid x_t)\mathrm{d}x_0 + \left|e^{Z(x_t)} - 1\right|\right\} \mathrm{d}x_t \\ &\overset{(b)}{\leq} \int_{x_t} p_{X_t}(x_t) \left\{\int_{x_0} (e^{\zeta(x_t, x_0)} - 1) e^{Z(x_t)} p_{X_0|X_t}(x_0 \mid x_t)\mathrm{d}x_0 + \left|e^{Z(x_t)} - 1\right|\right\} \mathrm{d}x_t \\ &= \int_{x_t, x_0} p_{X_t,X_0}(x_t, x_0) e^{\zeta(x_t, x_0) + Z(x_t)} \mathrm{d}x_t \mathrm{d}x_0 - \int_{x_t} p_{X_t}(x_t) e^{Z(x_t)} \mathrm{d}x_t + \int_{x_t} p_{X_t}(x_t) \left|e^{Z(x_t)} - 1\right| \mathrm{d}x_t.\end{aligned} \tag{87}$$

Note that $e^{\zeta(x_t, x_0) + Z(x_t)} = \mathcal{G}(x_t, x_0) = \frac{\widetilde{p}_{U_t|X_0}(u_t|x_0)}{p_{X_t|X_0}(x_t|x_0)} \det\left(\frac{\mathrm{d}u_t}{\mathrm{d}x_t}\right)$. As a result,

$$\int_{x_t, x_0} p_{X_t,X_0}(x_t, x_0) e^{\zeta(x_t, x_0) + Z(x_t)} \mathrm{d}x_t \mathrm{d}x_0 = \int_{x_t, x_0} \widetilde{p}_{U_t|X_0}(u_t|x_0) p(x_0) \mathrm{d}u_t \mathrm{d}x_0 = 1.$$

Substituting this into (87) gives

$$\begin{aligned}\mathcal{R}_{t-1} &\leq 1 - \int_{x_t} p_{X_t}(x_t) e^{Z(x_t)} \mathrm{d}x_t + \int_{x_t} p_{X_t}(x_t) \left|e^{Z(x_t)} - 1\right| \mathrm{d}x_t \\ &= 2\int_{x_t} p_{X_t}(x_t) \left(1 - e^{Z(x_t)}\right)_+ \mathrm{d}x_t \leq 2\mathbb{E}_{X_t}\left(-Z(X_t)\right)_+,\end{aligned} \tag{88}$$

where the last inequality holds by combining $1 - e^x \leq -x$ and the fact that the function $(\cdot)_+$ is non-decreasing.

C.5 Controlling the discretization error

From equation (88), we know $\mathcal{R}_{t-1} \leq \mathbb{E}_{X_t}\left[\left(-Z(X_t)\right)_+\right]$. By setting the square matrix Δ in Lemma 5 to 0, we can show that for any positive semi-definite matrix $A \in \mathbb{R}^{d\times d}$,

$$\log\det\left(I + A\right) \geq \mathrm{tr}(A) - \|A\|_{\mathrm{F}}^2. \tag{89}$$

Equipped with this result, we can derive the following inequality:

$$\begin{aligned}\mathbb{E}_{X_t}\left[(-Z(X_t))_+\right] &= \mathbb{E}_{X_t}\left[\left(\overline{\alpha}_t\eta_t'\mathrm{tr}\left(\mathrm{Cov}_{0|t}\right) - \log\det(I+\overline{\alpha}_t\eta_t^{(1)}\mathrm{Cov}_{0|t})\right)_+\right] \\ &\overset{(a)}{\leq} \mathbb{E}_{X_t}\left[\left(\overline{\alpha}_t\left(\eta_t^{(2)}-\eta_t^{(1)}\right)\mathrm{tr}\left(\mathrm{Cov}_{0|t}\right) + \overline{\alpha}_t^2\left(\eta_t^{(1)}\right)^2\left\|\mathrm{Cov}_{0|t}\right\|_{\mathrm{F}}^2\right)_+\right] \\ &= \overline{\alpha}_t\left(\eta_t^{(2)}-\eta_t^{(1)}\right)\mathbb{E}\left[\mathrm{tr}\left(\mathrm{Cov}_{0|t}\right)\right] + \overline{\alpha}_t^2\left(\eta_t^{(1)}\right)^2\mathbb{E}\left[\left\|\mathrm{Cov}_{0|t}\right\|_{\mathrm{F}}^2\right].\end{aligned} \tag{90}$$

Here, (a) holds by combining (89) and the fact that the function $(\cdot)_+$ is non-decreasing. Then we begin to deal with the coefficient $\overline{\alpha}_t\left(\eta_t^{(2)}-\eta_t^{(1)}\right)$ and $\overline{\alpha}_t^2\left(\eta_t^{(1)}\right)^2$. Combining the definition of $\eta_t^{(1)}$, $\eta_t^{(2)}$ and the relationship (26) results in

$$\begin{aligned}\overline{\alpha}_t\left(\eta_t^{(2)}-\eta_t^{(1)}\right) &= \overline{\alpha}_t\left(\frac{\eta_t}{(1-\overline{\alpha}_t)\overline{\sigma}_t^2} - \frac{\eta_t^2}{2(1-\overline{\alpha}_t)^2\overline{\sigma}_t^2} - \frac{\eta_t}{(1-\overline{\alpha}_t)(1-\overline{\alpha}_t-\eta_t)}\right) \\ &\leq \frac{\overline{\alpha}_t\eta_t}{1-\overline{\alpha}_t}\cdot\frac{1-\overline{\alpha}_t-\eta_t-\overline{\sigma}_t^2}{\overline{\sigma}_t^2(1-\overline{\alpha}_t-\eta_t)} = \frac{\overline{\alpha}_t\eta_t}{1-\overline{\alpha}_t}\cdot\frac{1-\overline{\alpha}_t-\eta_t-(1-\overline{\alpha}_t)\left(1-\frac{2\eta_t}{1-\overline{\alpha}_t}+\frac{\eta_t^2}{(1-\overline{\alpha}_t)^2}\right)}{(1-\overline{\alpha}_t-\eta_t)\overline{\sigma}_t^2} \\ &= \frac{\overline{\alpha}_t\eta_t}{1-\overline{\alpha}_t}\cdot\frac{\eta_t-\eta_t^2/(1-\overline{\alpha}_t)}{(1-\overline{\alpha}_t-\eta_t)\overline{\sigma}_t^2} \leq \frac{8\overline{\alpha}_t\eta_t^2}{(1-\overline{\alpha}_t)^3},\end{aligned}$$

where the last inequality holds by utilizing $\eta_t \leq \frac{1}{2}(1-\overline{\alpha}_t)$ and $\overline{\sigma}_t^2 = (1-\overline{\alpha}_t)\left(1-\frac{\eta_t}{1-\overline{\alpha}_t}\right)^2 \geq (1-\overline{\alpha}_t)\left(1-\frac{1}{2}\right)^2 = 4(1-\overline{\alpha}_t)$. Applying the same conclusion and arguments, we can also derive the following inequality:

$$\overline{\alpha}_t^2\left(\eta_t^{(1)}\right)^2 = \frac{\overline{\alpha}_t^2\eta_t^2}{(1-\overline{\alpha}_t)^2(1-\overline{\alpha}_t-\eta_t)^2} \leq \frac{4\overline{\alpha}_t^2\eta_t^2}{(1-\overline{\alpha}_t)^2(\alpha_t-\overline{\alpha}_t)^2} \leq \frac{4C^2(1-\alpha_t)^2}{(1-\overline{\alpha}_t)^2(\alpha_t-\overline{\alpha}_t)^2} = 4C^2\widetilde{\sigma}_t^4,$$

where $\widetilde{\sigma}_t^2 = \frac{\overline{\alpha}_t(1-\alpha_t)}{(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}$. According to Lemma 3, one has

$$\widetilde{\sigma}_t^4\mathbb{E}\left[\left\|\mathrm{Cov}_{0|t}\right\|_{\mathrm{F}}^2\right] \leq 3\widetilde{\sigma}_t^2\left\{\mathbb{E}\left[\mathrm{tr}(\mathrm{Cov}_{X_0|X_t})\right] - \mathbb{E}\left[\mathrm{tr}(\mathrm{Cov}_{X_0|X_{t-1}})\right]\right\} + \frac{3}{T^{10}}.$$

Back to (90), now we have

$$\mathbb{E}_{X_t}\left[\left(-Z(X_t)\right)_+\right] \leq \frac{8\overline{\alpha}_t\eta_t^2}{(1-\overline{\alpha}_t)^3}\mathbb{E}\left[\mathrm{tr}(\mathrm{Cov}_{0|t})\right] + C\widetilde{\sigma}_t^2\left\{\mathbb{E}\left[\mathrm{tr}(\mathrm{Cov}_{X_0|X_t})\right] - \mathbb{E}\left[\mathrm{tr}(\mathrm{Cov}_{X_0|X_{t-1}})\right]\right\} + \frac{C}{T^{10}}, \tag{91}$$

where C represents a new universal constant in the above inequality. Taking (79), (88) and (91) together results in

$$\begin{aligned}\mathsf{TV}(p_{X_{t-1}}, p_{\overline{Y}_{t-1}}) &\leq \mathcal{R}_{t-1} + \mathsf{TV}(p_{X_t}, p_{\overline{Y}_t}) \leq \mathsf{TV}(p_{X_t}, p_{\overline{Y}_t}) + \mathbb{E}_{X_t}\left[(-Z(X_t))_+\right] \\ &\leq \mathsf{TV}(p_{X_t}, p_{\overline{Y}_t}) + \left(C\widetilde{\sigma}_t^2 + \frac{\overline{\alpha}_t\eta_t^2}{(1-\overline{\alpha}_t)^3}\right)\mathbb{E}\left[\mathrm{tr}\left(\mathrm{Cov}_{0|t}\right)\right] - C\widetilde{\sigma}_t^2\mathbb{E}\left[\mathrm{tr}\left(\mathrm{Cov}_{0|t-1}\right)\right] + \frac{C}{T^{10}}.\end{aligned}$$

Applying the above result from 1 to T recursively, we arrive at

$$\begin{aligned}
\mathsf{TV}(p_{X_1}, p_{\overline{Y}_1}) &\le \mathsf{TV}(p_{X_T}, p_{Y_T^*}) + \sum_{t=2}^{T-1} \left(\frac{8\overline{\alpha}_t \eta_t^2}{(1-\overline{\alpha}_t)^3} + C(\widetilde{\sigma}_t^2 - \widetilde{\sigma}_{t+1}^2) \right) \mathbb{E}\left[\mathrm{tr}\left(\mathrm{Cov}_{0|t}\right)\right] \\
&\quad + \left(C\widetilde{\sigma}_t^2 + \frac{8\overline{\alpha}_t \eta_t^2}{(1-\overline{\alpha}_t)^3} \right) \mathbb{E}\left[\mathrm{tr}\left(\mathrm{Cov}_{0|T}\right)\right] + \frac{C}{T^9} \\
&\overset{(a)}{\le} \mathsf{TV}(p_{X_T}, p_{\overline{Y}_T}) + \left(\frac{C\log T}{T}\right)^2 \sum_{t=1}^{T-1} \frac{\overline{\alpha}_t}{1-\overline{\alpha}_t} \mathbb{E}\left[\mathrm{tr}\left(\mathrm{Cov}_{0|t}\right)\right] \\
&\quad + \frac{C\log T}{T} \frac{\overline{\alpha}_T}{1-\overline{\alpha}_T} \mathbb{E}\left[\mathrm{tr}\left(\mathrm{Cov}_{0|T}\right)\right] + \frac{1}{T^9} \\
&\overset{(b)}{\le} \mathsf{TV}(p_{X_T}, p_{\overline{Y}_T}) + C_3 k T \log T \left(\frac{C\log T}{T}\right)^2 + \frac{C_3 C k \log^2 T}{T} + \frac{C}{T^9} \\
&\le \frac{1}{T^{10}} + C^2 C_3 \frac{k\log^3 T}{T} + \frac{C}{T^9} \le C_8 \frac{k \log^3 T}{T}.
\end{aligned} \tag{92}$$

Here, (a) applies Lemma 4 and the basic property (46), (b) makes use of the moment inequality (45) with $l = 2$, whereas the penultimate inequality results from Li and Yan (2024a, Lemma 10).

C.6 Controlling the term related to score estimation errors

In this subsection, we bound the TV distance between p_{Y_1} and $p_{\overline{Y}_1}$. Observe that

$$\begin{aligned}
\mathsf{TV}(p_{Y_1}, p_{\overline{Y}_1}) &= \int_{\mathbb{R}^d} \left(p_{\overline{Y}_1}(x) - p_{Y_1}(x)\right) \mathbb{1}\{p_{\overline{Y}_1}(x) > p_{Y_1}(x)\} \mathrm{d}x + \mathbb{P}(\overline{Y}_1 = \infty) \\
&\overset{(a)}{\le} \int_{\mathbb{R}^d} \left(p_{\overline{Y}_1}(x) - p_{\widehat{Y}_1}(x)\right) \mathbb{1}\{p_{\overline{Y}_1}(x) > p_{\widehat{Y}_1}(x)\} \mathrm{d}x + \mathbb{P}(\overline{Y}_1 = \infty) \\
&\overset{(b)}{\le} \mathsf{TV}(p_{\overline{Y}_1}, p_{\widehat{Y}_1}) + \mathsf{TV}(p_{X_1}, p_{\overline{Y}_1}) \overset{(c)}{\le} \sqrt{\mathsf{KL}(p_{\overline{Y}_1} \parallel p_{\widehat{Y}_1})} + C\frac{k\log^3 T}{T},
\end{aligned} \tag{93}$$

where (a) holds due to (78), (b) follows since $\mathbb{P}(\overline{Y}_0 = \infty) \le \mathsf{TV}(X_0, \overline{Y}_0)$, and (c) invokes Pinsker's inequality and (92). Regarding the KL divergence term, it follows from the data processing inequality that

$$\begin{aligned}
&\mathsf{KL}(p_{\overline{Y}_1} \parallel p_{\widehat{Y}_1}) \le \mathsf{KL}\left(p_{\overline{Y}_1, \overline{Y}_1^-, \cdots, \overline{Y}_T, \overline{Y}_T^-} \parallel p_{\widehat{Y}_1, \widehat{Y}_1^-, \cdots, \widehat{Y}_T, \widehat{Y}_T^-}\right) \\
&\overset{(a)}{=} \mathsf{KL}(p_{\overline{Y}_T^-} \parallel p_{\widehat{Y}_T^-}) + \sum_{t=2}^{T} \mathbb{E}_{x_t \sim p_{\overline{Y}_t}} \left[\mathsf{KL}\left(p_{\overline{Y}_{t-1}^- | \overline{Y}_t = x_t} \parallel p_{\widehat{Y}_{t-1}^- | \widehat{Y}_t = x_t}\right)\right] \\
&+ \sum_{t=2}^{T} \mathbb{E}_{x_t \sim p_{\overline{Y}_t^-}} \left[\mathsf{KL}\left(p_{\overline{Y}_{t-1} | \overline{Y}_t^- = x_t} \parallel p_{\widehat{Y}_{t-1} | \widehat{Y}_t^- = x_t}\right)\right] \\
&\overset{(b)}{=} \sum_{t=2}^{T} \mathbb{E}_{x_t \sim p_{\overline{Y}_t}} \left[\mathsf{KL}\left(p_{\overline{Y}_{t-1}^- | \overline{Y}_t = x_t} \parallel p_{\widehat{Y}_{t-1}^- | \widehat{Y}_t = x_t}\right)\right] \overset{(c)}{=} \sum_{t=2}^{T} \mathbb{E}_{x_t \sim p_{\overline{Y}_t}} \left[\mathsf{KL}\left(p_{Y_{t-1}^* | Y_t^* = x_t} \parallel p_{Y_{t-1} | Y_t = x_t}\right)\right].
\end{aligned}$$

Here, (a) follows from the chain rule of KL divergence, (b) holds since the conditional distributions of $\widehat{Y}_t$ given $\widehat{Y}_t^- = x$ and $\overline{Y}_t$ given $\overline{Y}_t^- = x$ are identical, and (c) arises from the construction of the two sequences $\overline{Y}_{t-1}^- \mid \overline{Y}_t$ and $\widehat{Y}_{t-1}^- \mid \widehat{Y}_t$.

Recall that $Y_{t-1}^* \mid Y_t^* = x_t$ and $Y_{t-1} \mid Y_t = x_t$ are two Gaussian distributions with different means but the same covariance. More precisely,

$$Y_{t-1}^* \mid Y_t^* = x_t \sim \mathcal{N}\left(\frac{x_t + \eta_t s_t^*(x_t)}{\sqrt{\alpha_t}}, \sigma_t^2 I_d\right), \quad Y_{t-1} \mid Y_t = x_t \sim \mathcal{N}\left(\frac{x_t + \eta_t s_t(x_t)}{\sqrt{\alpha_t}}, \sigma_t^2 I_d\right)$$

with $\eta_t,\ \sigma_t^2$ satisfying $\eta_t^2 \le C(1-\alpha_t)\sigma_t^2$. Further, the KL divergence between two Gaussian measures can be computed in closed form, i.e.,

$$\begin{aligned}
&\mathsf{KL}\left(p_{Y^*_{t-1}|Y^*_t}(\cdot|x_t) \parallel p_{Y_{t-1}|Y_t}(\cdot|x_t)\right) = \mathsf{KL}\left(\mathcal{N}\left(\frac{x_t+\eta_t s_t^*(x_t)}{\sqrt{\alpha_t}}, \sigma_t^2 I_d\right) \,\middle\|\, \mathcal{N}\left(\frac{x_t+\eta_t s_t(x_t)}{\sqrt{\alpha_t}}, \sigma_t^2 I_d\right)\right) \\
&= \frac{\eta_t^2/\alpha_t}{2\sigma_t^2}\left\|s_t(x_t)-s_t^*(x_t)\right\|_2^2 = \frac{\eta_t^2/\alpha_t}{2\sigma_t^2}\left\|\varepsilon_t^{\mathsf{sc}}(x_t)\right\|_2^2 = \frac{C(1-\alpha_t)}{2\alpha_t}\left\|\varepsilon_t^{\mathsf{sc}}(x_t)\right\|_2^2 \le C(1-\alpha_t)\left\|\varepsilon_t^{\mathsf{sc}}(x_t)\right\|_2^2.
\end{aligned}$$

Therefore, we can calculate that

$$\begin{aligned}
\mathsf{KL}(p_{\overline{Y}_1} \parallel p_{\widehat{Y}_1}) &\le \sum_{t=2}^{T} \mathbb{E}_{x_t\sim p_{\overline{Y}_t}}\left[\mathsf{KL}\left(p_{Y^*_{t-1}|Y^*_t=x_t} \,\middle\|\, p_{Y_{t-1}|Y_t=x_t}\right)\right] \\
&\le \sum_{t=2}^{T} C(1-\alpha_t)\mathbb{E}_{x_t\sim p_{\overline{Y}_t}}\left[\left\|\varepsilon_t^{\mathsf{sc}}(x_t)\right\|_2^2\right] \overset{(a)}{\le} \sum_{t=2}^{T} C(1-\alpha_t)\mathbb{E}_{x_t\sim p_{X_t}}\left[\left\|\varepsilon_t^{\mathsf{sc}}(x_t)\right\|_2^2\right] \\
&\le \frac{c_1 C\log T}{T}\sum_{t=2}^{T}\varepsilon_{\mathsf{score},t}^2 \le c_1 C(\log T)\varepsilon_{\mathsf{score}}^2,
\end{aligned} \tag{94}$$

where (a) follows from (77). Combine (92), (93) and (94) to reach

$$\begin{aligned}
\mathsf{TV}(p_{X_1}, p_{Y_1}) &\le \mathsf{TV}(p_{X_1}, p_{\overline{Y}_1}) + \mathsf{TV}(p_{\overline{Y}_1}, p_{Y_1}) \\
&\le C\frac{k\log^3 T}{T} + C\frac{k\log^3 T}{T} + \sqrt{\mathsf{KL}(p_{\overline{Y}_1} \parallel p_{\widehat{Y}_1})} \\
&\le 2C\frac{k\log^3 T}{T} + \sqrt{c_1\varepsilon_{\mathsf{score}}^2\log T} = C\frac{k\log^3 T}{T} + \sqrt{c_1 C\log T}\,\varepsilon_{\mathsf{score}},
\end{aligned}$$

thereby concluding the proof of Theorem 3.

D Equivalence between relation (26) and Song et al. (2020, Eq. (12))

Recall that $\epsilon_t^{\mathsf{noise}}(Y_t) = -\sqrt{1-\overline{\alpha}_t}\,s_t(Y_t)$. Substituting this expression into (29) (i.e., Song et al. (2020, Eq. (12))) results in:

$$Y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\Big(Y_t + (1-\overline{\alpha}_t)s_t(Y_t) - \sqrt{(1-\overline{\alpha}_t)(\alpha_t-\overline{\alpha}_t-\alpha_t\varsigma_t^2)}\,s_t(Y_t) + \sqrt{\alpha_t}\varsigma_t Z_t\Big).$$

By taking

$$\eta_t^{\mathsf{ddpm}} = (1-\overline{\alpha}_t) - \sqrt{(1-\overline{\alpha}_t)(\alpha_t-\overline{\alpha}_t-\alpha_t\varsigma_t^2)} \qquad \text{and} \qquad \sigma_t^{\mathsf{ddpm}} = \sqrt{\alpha_t}\varsigma_t,$$

we can derive

$$\begin{aligned}
(1-\overline{\alpha}_t)\left(1-\frac{\eta_t^{\mathsf{ddpm}}}{1-\overline{\alpha}_t}\right)^2 &= (1-\overline{\alpha}_t)\left(1-\frac{(1-\overline{\alpha}_t)-\sqrt{(1-\overline{\alpha}_t)\big(\alpha_t-\overline{\alpha}_t-\big(\sigma_t^{\mathsf{ddpm}}\big)^2\big)}}{1-\overline{\alpha}_t}\right)^2 \\
&= (1-\overline{\alpha}_t)\left(\sqrt{\frac{\alpha_t-\overline{\alpha}_t-\big(\sigma_t^{\mathsf{ddpm}}\big)^2}{1-\overline{\alpha}_t}}\right)^2 = \alpha_t-\overline{\alpha}_t-\big(\sigma_t^{\mathsf{ddpm}}\big)^2,
\end{aligned}$$

which is precisely the relation in (26).

E Proofs about reverse-time differential equations

E.1 Generalized reverse-time differential equations

We formally state the time-reversal property of the generalized reverse-time differential equation introduced in Section 2.

Proposition 3 *Suppose the generalized reverse-time differential equation*

$$\mathrm{d}Y_t = \big(Y_t + \big(1+\xi(T-t)\big)s^*_{T-t}(Y_t)\big)\,\beta(T-t)\mathrm{d}t + \sqrt{2\xi(T-t)\beta(T-t)}\,\mathrm{d}W_t, \qquad t\in[0,T] \tag{95}$$

has a unique strong solution, where (W_t) represents a standard Brownian motion in $\mathbb{R}^d$. Then under the boundary condition $Y_0 \overset{\mathrm{d}}{=} X_T$, it satisfies $Y_{T-t} \overset{\mathrm{d}}{=} X_t$ for all $0\le t\le T$.

Proof of Proposition 3. Recall that the continuous-time forward process is given by

$$\mathrm{d}X_t = -\beta(t)X_t\mathrm{d}t + \sqrt{2\beta(t)}\,\mathrm{d}B_t,$$

with (B_t) a standard Brownian motion in $\mathbb{R}^d$. Denote by $p_X(x,t)$ the probability density of X_t at point x w.r.t. the Lebesgue measure in $\mathbb{R}^d$. In the following proof, we use ∇ (resp. $\nabla\cdot$) to be the gradient (resp. divergence) operator taken w.r.t. the first argument (i.e., x) of the function, and denote by Δ the corresponding Laplace operator. The Fokker-Planck equation then tells us that

$$\begin{aligned}\frac{\partial}{\partial t}p_X(x,t) &= \nabla\cdot\big(x\beta(t)p_X(x,t)\big) + \frac{1}{2}\Delta\big(2\beta(t)p_X(x,t)\big)\\ &= \beta(t)\nabla\cdot\big(xp_X(x,t)\big) + \beta(t)\Delta\big(p_X(x,t)\big).\end{aligned} \tag{96}$$

Similarly, denoting by $p_Y(x,t)$ the probability density of Y_t at point x w.r.t. the Lebesgue measure in $\mathbb{R}^d$, then we can apply the Fokker-Planck equation once again to obtain

$$\begin{aligned}\frac{\partial}{\partial t}p_Y(x,t) &= -\nabla\cdot\Big(\Big(x+\big(1+\xi(T-t)\big)s^*_{T-t}(x)\Big)\beta(T-t)p_Y(x,t)\Big) + \frac{1}{2}\Delta\big(2\xi(T-t)\beta(T-t)p_Y(x,t)\big)\\ &= -\Big\langle x+\big(1+\xi(T-t)\big)s^*_{T-t}(x),\,\beta(T-t)\nabla p_Y(x,t)\Big\rangle\\ &\qquad - \operatorname{tr}\big(I_d + \big(1+\xi(T-t)\big)\nabla s^*_{T-t}(x)\big)\,\beta(T-t)p_Y(x,t) + \xi(T-t)\beta(T-t)\Delta\big(p_Y(x,t)\big).\end{aligned} \tag{97}$$

Recall that our goal is to show that X_t and Y_{T-t} have the same marginal distributions, i.e., $p_X(x,t) = p_Y(x,T-t)$, or equivalently, $p_X(x,T-t) = p_Y(x,t)$. Since the generalized differential equation (95) is assumed to have a unique strong solution, the induced Fokker-Planck equation has a unique strong solution. From the assumption $Y_0 \overset{\mathrm{d}}{=} X_T$, we know that $p_X(x,T) = p_Y(x,0)$, and hence it suffices to show that $p_X(x,T-t)$ is a solution of the partial differential equation (PDE) (97). It is readily seen from PDE (96) that

$$\frac{\partial}{\partial t}p_X(x,T-t) = -\beta(T-t)\nabla\cdot\big(xp_X(x,T-t)\big) - \beta(T-t)\Delta\big(p_X(x,T-t)\big). \tag{98}$$

Replacing $p_Y(x,t)$ with $p_X(x,T-t)$ on the right-hand side of PDE (97), one can derive

$$\begin{aligned}&-\Big\langle x+\big(1+\xi(T-t)\big)s^*_{T-t}(x),\,\beta(T-t)\nabla p_X(x,T-t)\Big\rangle - \operatorname{tr}(I_d)\beta(T-t)p_X(x,T-t)\\ &\qquad - \big(1+\xi(T-t)\big)\beta(T-t)\Delta\big(\log(p_X(x,T-t))\big)p_X(x,T-t) + \xi(T-t)\beta(T-t)\Delta\big(p_X(x,T-t)\big)\\ &= -\beta(T-t)\Big\langle x+\big(1+\xi(T-t)\big)\frac{\nabla p_X(x,T-t)}{p_X(x,T-t)},\nabla p_X(x,T-t)\Big\rangle - d\beta(T-t)p_X(x,T-t)\\ &\qquad + \big(1+\xi(T-t)\big)\beta(T-t)\frac{\|\nabla p_X(x,T-t)\|_2^2}{p_X(x,T-t)} + \Big(-\big(1+\xi(T-t)\big)+\xi(T-t)\Big)\beta(T-t)\Delta\big(p_X(x,T-t)\big)\end{aligned}$$

$$
\begin{aligned}
&= -\beta(T-t)\big\langle x, \nabla p_X(x,T-t)\big\rangle - d\beta(T-t)p_X(x,T-t) - \beta(T-t)\Delta\big(p_X(x,T-t)\big) \\
&= -\beta(T-t)\nabla\cdot\big(xp_X(x,T-t)\big) - \beta(T-t)\Delta\big(p_X(x,T-t)\big) \\
&= \frac{\partial}{\partial t}p_X(x,T-t), \qquad (99)
\end{aligned}
$$

where we invoke PDE (98) in the last line. Eqn. (99) reveals that $p_X(x,T-t)$ is a strong solution of PDE (97), which is equivalent to $p_X(x,T-t) = p_Y(x,t)$.

E.2 Proof of Proposition 2

We prove this result by explicitly solving SDE (33) when $t \in [t_n, t_{n+1})$. To begin with, we make the observation that: under the time transformation

$$
t \;\to\; t' = \int_0^t \beta(s)\mathrm{d}s, \qquad (100)
$$

SDE (33) can be rewritten as

$$
\mathrm{d}\widetilde{Y}_{t'} = \left(-\frac{\xi(T-t'_n)+\overline{\alpha}_{T-t'}}{1-\overline{\alpha}_{T-t'}}\widetilde{Y}_{t'} + \frac{(1+\xi(T-t'_n))\sqrt{\overline{\alpha}_{T-t'}}}{1-\overline{\alpha}_{T-t'}}\mu_{T-t'_n}(\widetilde{Y}_{t'_n})\right)\mathrm{d}t' + \sqrt{2\xi(T-t'_n)}\,\mathrm{d}W_{t'}
$$

for $t \in [t'_n, t'_{n+1})$, where t'_n and t'_{n+1} are the images of t_n and t_{n+1} under the transformation (100). Note that this transformed SDE has the same form as SDE (33) when $\beta(t) = 1$ for $t \in [t'_n, t'_{n+1})$. Thus, without loss of generality, it suffices to assume $\beta(t) = 1$ for all $t \in [0,T]$ and solve SDE (33). Under this assumption, we can simplify

$$
\overline{\alpha}_t = \exp\left(-2\int_0^t \beta(s)\mathrm{d}s\right) = e^{-2t}. \qquad (101)
$$

Recall that SDE (33) with $\xi(T-t_n) = \xi > 0$ and $\beta(t) = 1$ can be written as

$$
\mathrm{d}\widetilde{Y}_t = \left(-\frac{\xi+\overline{\alpha}_{T-t}}{1-\overline{\alpha}_{T-t}}\widetilde{Y}_t + \frac{(1+\xi)\sqrt{\overline{\alpha}_{T-t}}}{1-\overline{\alpha}_{T-t}}\mu_{T-t_n}(\widetilde{Y}_{t_n})\right)\mathrm{d}t + \sqrt{2\xi}\,\mathrm{d}W_t. \qquad (102)
$$

To solve SDE (102), we find it convenient to introduce the following function

$$
f(t) = \frac{e^{-\xi(T-t)}}{(1-e^{-2(T-t)})^{\frac{1+\xi}{2}}}.
$$

It follows from Itô's formula that

$$
\begin{aligned}
\mathrm{d}\left(f(t)\widetilde{Y}_t\right) &= \left(\left(f'(t) - \frac{\xi+\overline{\alpha}_{T-t}}{1-\overline{\alpha}_{T-t}}f(t)\right)\widetilde{Y}_t + \frac{(1+\xi)\sqrt{\overline{\alpha}_{T-t}}}{1-\overline{\alpha}_{T-t}}f(t)\mu_{T-t_n}(\widetilde{Y}_{t_n})\right)\mathrm{d}t + \sqrt{2\xi}f(t)\,\mathrm{d}W_t \\
&= \frac{(1+\xi)\sqrt{\overline{\alpha}_{T-t}}}{1-\overline{\alpha}_{T-t}}f(t)\mu_{T-t_n}(\widetilde{Y}_{t_n})\mathrm{d}t + \sqrt{2\xi}f(t)\,\mathrm{d}W_t.
\end{aligned}
$$

Integrating both sides of the above display from t_n to t_{n+1}, we obtain

$$
f(t_{n+1})\widetilde{Y}_{t_{n+1}} - f(t_n)\widetilde{Y}_{t_n} = \int_{t_n}^{t_{n+1}} \frac{(1+\xi)\sqrt{\overline{\alpha}_{T-t}}}{1-\overline{\alpha}_{T-t}}f(t)\mu_{T-t_n}(\widetilde{Y}_{t_n})\mathrm{d}t + \int_{t_n}^{t_{n+1}} \sqrt{2\xi}f(t)\,\mathrm{d}W_t.
$$

From Itô's isometry property of the Brownian motion, we can write, for each $0 \le n \le T-1$,

$$
\int_{t_n}^{t_{n+1}} \sqrt{2\xi}f(t)\,\mathrm{d}W_t = \left(\int_{t_n}^{t_{n+1}} 2\xi\big(f(t)\big)^2\mathrm{d}t\right)^{1/2}\widetilde{Z}_n
$$

for some Gaussian vector $\widetilde{Z}_n \sim \mathcal{N}(0, I_d)$, where $\{\widetilde{Z}_n\}_{n=0,\ldots,T-1}$ are statistically independent. Consequently,

$$f(t_{n+1})\widetilde{Y}_{t_{n+1}} = f(t_n)\widetilde{Y}_{t_n} + \underbrace{\int_{t_n}^{t_{n+1}} \frac{(1+\xi)\sqrt{\overline{\alpha}_{T-t}}}{1-\overline{\alpha}_{T-t}} f(t)\mathrm{d}t}_{=:A_n} \cdot \mu_{T-t_n}(\widetilde{Y}_{t_n}) + \left(\underbrace{\int_{t_n}^{t_{n+1}} 2\xi\big(f(t)\big)^2 \mathrm{d}t}_{=:B_n}\right)^{1/2} \cdot \widetilde{Z}_n.$$

Taking this together with the definition (31b) of μ_t, we can express the update rule induced by SDE (102) as

$$\widetilde{Y}_{t_{n+1}} = \frac{f(t_n) + A_n/\sqrt{\overline{\alpha}_{T-t_n}}}{f(t_{n+1})}\widetilde{Y}_{t_n} + \frac{1-\overline{\alpha}_{T-t_n}}{\sqrt{\overline{\alpha}_{T-t_n}}} \cdot \frac{A_n}{f(t_{n+1})} s_{T-t_n}(\widetilde{Y}_{t_n}) + \frac{B_n}{f(t_{n+1})}\widetilde{Z}_n. \tag{103}$$

To simplify the notation, we define

$$\gamma_n = e^{-(T-t_n)}.$$

The terms A_n and B_n can be explicitly calculated as follows:

$$\begin{aligned} A_n &= \int_{t_n}^{t_{n+1}} \frac{(1+\xi)\sqrt{\overline{\alpha}_{T-t}}}{1-\overline{\alpha}_{T-t}} f(t)\mathrm{d}t \\ &= \int_{t_n}^{t_{n+1}} (1+\xi)\, \frac{e^{-(1+\xi)(T-t)}}{\big(1-e^{-2(T-t)}\big)^{(3+\xi)/2}}\, \mathrm{d}t \\ &= \frac{e^{-(1+\xi)(T-t)}}{\big(1-e^{-2(T-t)}\big)^{(1+\xi)/2}}\Bigg|_{t_n}^{t_{n+1}} = \frac{\gamma_{n+1}^{\xi+1}}{(1-\gamma_{n+1}^2)^{\frac{1+\xi}{2}}} - \frac{\gamma_n^{\xi+1}}{(1-\gamma_n^2)^{\frac{1+\xi}{2}}}, \end{aligned} \tag{104}$$

where we have applied (101). Through similar calculation, we can reach

$$\begin{aligned} B_n &= \int_{t_n}^{t_{n+1}} 2\xi\big(f(t)\big)^2 \mathrm{d}t \\ &= \int_{t_n}^{t_{n+1}} 2\xi \frac{e^{-2\xi(T-t)}}{\big(1-e^{-2(T-t)}\big)^{1+\xi}}\, \mathrm{d}t \\ &= \frac{e^{-2\xi(T-t)}}{\big(1-e^{-2(T-t)}\big)^{\xi}}\Bigg|_{t_n}^{t_{n+1}} = \frac{\gamma_{n+1}^{2\xi}}{(1-\gamma_{n+1}^2)^{\xi}} - \frac{\gamma_n^{2\xi}}{(1-\gamma_n^2)^{\xi}}. \end{aligned} \tag{105}$$

Substituting (104) and (105) into (103) and comparing the coefficients with the DDPM update rule (4), we obtain

$$\alpha_{t_n} = \frac{f(t_n) + A_n/\sqrt{\overline{\alpha}_{T-t_n}}}{f(t_{n+1})} = \left(\frac{\gamma_n}{\gamma_{n+1}}\right)^2 = e^{-2(t_{n+1}-t_n)}, \tag{106}$$

which coincides with our choice of $\overline{\alpha}_t$, i.e.,

$$\alpha_{t_n} = e^{-2(t_{n+1}-t_n)} = \frac{\overline{\alpha}_{t_{n+1}}}{\overline{\alpha}_{t_n}}.$$

Additionally, we can easily verify that

$$\begin{aligned} \eta_{t_n}^{\mathsf{ddpm}} &= \sqrt{\alpha_{t_n}} \cdot \frac{1-\gamma_n^2}{\gamma_n} \cdot \frac{A_n}{f(t_{n+1})} = \frac{1-\gamma_n^2}{\gamma_{n+1}} \cdot \frac{1}{f(t_{n+1})} \cdot \left(\frac{\gamma_{n+1}^{\xi+1}}{(1-\gamma_{n+1}^2)^{\frac{1+\xi}{2}}} - \frac{\gamma_n^{\xi+1}}{(1-\gamma_n^2)^{\frac{1+\xi}{2}}}\right), \\ \sigma_{t_n}^{\mathsf{ddpm}} &= \sqrt{\alpha_{t_n}} \cdot \frac{B_n^{1/2}}{f(t_{n+1})} = \frac{\gamma_n}{\gamma_{n+1} f(t_{n+1})} \cdot \left(\frac{\gamma_{n+1}^{2\xi}}{(1-\gamma_{n+1}^2)^{\xi}} - \frac{\gamma_n^{2\xi}}{(1-\gamma_n^2)^{\xi}}\right)^{1/2}. \end{aligned}$$

We are now ready to show that the relation (26)

$$(1-\overline{\alpha}_{t_n})\left(1-\frac{\eta_{t_n}^{\mathsf{ddpm}}}{1-\overline{\alpha}_{t_n}}\right)^2 = \alpha_{t_n} - \overline{\alpha}_{t_n} - \left(\sigma_{t_n}^{\mathsf{ddpm}}\right)^2$$

is satisfied by this solution for all n. Towards this end, calculate the left-hand side above as:

$$\begin{aligned}
(1-\overline{\alpha}_{t_n})\left(1-\frac{\eta_{t_n}^{\mathsf{ddpm}}}{1-\overline{\alpha}_{t_n}}\right)^2 &= (1-\gamma_n^2)\left(1-\frac{1}{\gamma_{n+1}f(t_{n+1})}\cdot\left(\frac{\gamma_{n+1}^{\xi+1}}{(1-\gamma_{n+1}^2)^{\frac{1+\xi}{2}}}-\frac{\gamma_n^{\xi+1}}{(1-\gamma_n^2)^{\frac{1+\xi}{2}}}\right)\right)^2 \\
&= (1-\gamma_n^2)\left(\frac{\gamma_n^{\xi+1}}{\gamma_{n+1}^{\xi+1}}\cdot\frac{(1-\gamma_{n+1}^2)^{\frac{1+\xi}{2}}}{(1-\gamma_n^2)^{\frac{1+\xi}{2}}}\right)^2 \\
&= \frac{\gamma_n^{2(\xi+1)}}{\gamma_{n+1}^{2(\xi+1)}}\cdot\frac{(1-\gamma_{n+1}^2)^{1+\xi}}{(1-\gamma_n^2)^{\xi}} \\
&= \frac{\gamma_n^2}{\gamma_{n+1}^2}-\gamma_n^2-(1-\gamma_{n+1}^2)\left(\frac{\gamma_n^2}{\gamma_{n+1}^2}-\frac{\gamma_n^{2(\xi+1)}}{\gamma_{n+1}^{2(\xi+1)}}\cdot\frac{(1-\gamma_{n+1}^2)^{\xi}}{(1-\gamma_n^2)^{\xi}}\right) \\
&= \alpha_{t_n}-\overline{\alpha}_{t_n}-\left(\sigma_{t_n}^{\mathsf{ddpm}}\right)^2.
\end{aligned}$$

Thus, setting $t_n = T - n$ for $n = 0, 1, \ldots, T$ exactly recovers the relation (26).

E.3 Proof of Proposition 1

Proposition 1 can be regarded as a corollary of Proposition 2 in the following sense: if we set $\xi(T-t_n) = 0$ for all $n = 0, 1, \ldots, T-1$, then SDE (33) degenerates to ODE (32). In addition, the whole proof of Proposition 2 in Appendix E.2 works for $\xi(T - t_n) = 0$. Thus, the proof of Proposition 1 can be directly completed by repeating the proof arguments in Appendix E.2.

F Proof of the lower bound in Theorem 4

Let $X_0 \sim \mathcal{N}\left(0, \begin{bmatrix} I_k & \\ & 0 \end{bmatrix}\right)$, then it follows from (7) that

$$X_t = \sqrt{\overline{\alpha}_t}X_0 + \sqrt{1-\overline{\alpha}_t}\,\overline{W}_t \sim \mathcal{N}\left(0, \begin{bmatrix} I_k & \\ & (1-\overline{\alpha}_t)I_{d-k} \end{bmatrix}\right). \tag{107}$$

It is then easily seen that

$$s_t^*(x) = -\begin{bmatrix} I_k & \\ & (1-\overline{\alpha}_t)I_{d-k} \end{bmatrix}^{-1} x = \begin{bmatrix} I_k & \\ & \frac{1}{1-\overline{\alpha}_t}I_{d-k} \end{bmatrix} x.$$

As a result, the mapping Φ_t^* admits a closed-form expression as follows

$$\Phi_t^*(x,z) = \frac{1}{\sqrt{\alpha}_t}\left(x + \eta_t s_t^*(x) + \sigma_t z\right) = \frac{1}{\sqrt{\alpha}_t}(A_{\eta_t}x + \sigma_t z)$$

where

$$A_{\eta_t} := \begin{bmatrix} (1-\eta_t)I_k & \\ & \left(1-\frac{\eta_t}{1-\overline{\alpha}_t}\right)I_{d-k} \end{bmatrix}.$$

These properties taken together further imply that

$$\Phi_t^*(X_t, Z_t) \sim \mathcal{N}\left(0, \frac{1}{\alpha_t}A_{\eta_t}\begin{bmatrix} I_k & \\ & (1-\overline{\alpha}_t)I_{d-k} \end{bmatrix}A_{\eta_t} + \frac{\sigma_t^2}{\alpha_t}I_d\right),$$

or equivalently,

$$\Phi_t^*(X_t, Z_t) \sim \mathcal{N}\left(0, \begin{bmatrix} \frac{(1-\eta_t)^2}{\alpha_t} I_k & \\ & \frac{1-\overline{\alpha}_t}{\alpha_t}\left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right)^2 I_{d-k} \end{bmatrix} + \frac{\sigma_t^2}{\alpha_t} I_d\right). \tag{108}$$

Armed with the above basic properties, we can proceed to derive the advertised lower bound. Towards this end, we resort to the following result concerning the TV distance between two multivariate Gaussians with the same mean, whose proof can be found in Devroye et al. (2018).

Lemma 7 (TV distance between Gaussians with the same mean) *Consider any $\mu \in \mathbb{R}^d$, and any positive semidefinite matrices $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d\times d}$. Then it holds that*

$$\frac{1}{100} < \frac{\mathsf{TV}\big(\mathcal{N}(\mu, \Sigma_1),\ \mathcal{N}(\mu, \Sigma_2)\big)}{\min\left\{1,\ \left\|\Sigma_1^{-1}\Sigma_2 - I\right\|_{\mathrm{F}}\right\}} \le \frac{3}{2}.$$

Recall from (107) that

$$X_{t-1} \sim \mathcal{N}\left(0, \begin{bmatrix} I_k & \\ & (1-\overline{\alpha}_{t-1}) I_{d-k} \end{bmatrix}\right).$$

With this and (108) in mind, we take

$$\Sigma_1 = \begin{bmatrix} I_k & \\ & (1-\overline{\alpha}_{t-1}) I_{d-k} \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} \frac{(1-\eta_t)^2}{\alpha_t} I_k & \\ & \frac{1-\overline{\alpha}_t}{\alpha_t}\left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right)^2 I_{d-k} \end{bmatrix} + \frac{\sigma_t^2}{\alpha_t} I_d,$$

which satisfy

$$\Sigma_1^{-1}\Sigma_2 = \begin{bmatrix} \frac{(1-\eta_t)^2 + \sigma_t^2}{\alpha_t} I_k & \\ & \left(\frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}\left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right)^2 + \frac{\sigma_t^2}{\alpha_t - \overline{\alpha}_t}\right) I_{d-k} \end{bmatrix}.$$

Invoke Lemma 7 to arrive at the following lower bound:

$$\begin{aligned}
\mathsf{TV}\big(X_{t-1},\ \Phi_t^*(X_t)\big) &\ge \frac{1}{100}\min\left\{1,\ \left\|\Sigma_1^{-1}\Sigma_2 - I\right\|_{\mathrm{F}}\right\} \\
&= \frac{1}{100}\min\left\{1,\ \sqrt{k\left(\frac{(1-\eta_t)^2 + \sigma_t^2}{\alpha_t} - 1\right)^2 + (d-k)\left(\frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}\left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right)^2 + \frac{\sigma_t^2}{\alpha_t - \overline{\alpha}_t} - 1\right)^2}\right\} \\
&\ge \frac{1}{100}\min\left\{1,\ \sqrt{\frac{d}{2}\left(\frac{1-\overline{\alpha}_t}{\alpha_t - \overline{\alpha}_t}\left(1 - \frac{\eta_t}{1-\overline{\alpha}_t}\right)^2 + \frac{\sigma_t^2}{\alpha_t - \overline{\alpha}_t} - 1\right)^2}\right\},
\end{aligned}$$

where the last line follows from our assumption that $d \ge 2k$. This concludes the proof.

G Auxiliary lemmas and related proofs

G.1 Proof of Lemma 1

Recall that $V_\alpha = \sqrt{\alpha}V_1 + \sqrt{1-\alpha}W$, where $V_1 \sim p_{\mathsf{data}}$ and $W \sim \mathcal{N}(0, I_d)$. From this, we can derive that

$$\mathbb{P}(V_\alpha \notin \mathcal{T}_\alpha) = \mathbb{P}(\sqrt{\alpha}V_1 + \sqrt{1-\alpha}W \notin \mathcal{T}_\alpha) \le \mathbb{P}\left(\left\{V_1 \notin \bigcup_{i\in\mathcal{I}} \mathcal{B}_i\right\} \cup \{W \notin \mathcal{G}\}\right)$$

$$\leq \sum_{j\in[N_{\epsilon_0}]\setminus\mathcal{I}} \mathbb{P}(V_1 \in \mathcal{B}_j) + \mathbb{P}(W \notin \mathcal{G}).$$

According to the definition (37) of $\mathcal{I}$, we know that, for any $j \notin \mathcal{I}$,

$$\mathbb{P}(V_1 \in \mathcal{B}_j) \leq \exp\{-C_1 k \log T\},$$

with $C_1 > 0$ a universal constant. Taking this with Assumption 1 yields

$$\sum_{j\in[N_{\epsilon_0}]\setminus\mathcal{I}} \mathbb{P}(V_1 \in \mathcal{B}_j) \leq N_{\epsilon_0} \exp\{-C_1 k \log T\}$$
$$\leq \exp\left\{C_{\mathsf{cover}} k \log T - C_1 k \log T\right\} \leq \exp\left\{-\frac{3}{8} C_1 k \log T\right\},$$

where the last inequality holds as long as $C_1 \geq 16 C_{\mathsf{cover}}$.

In addition, we can establish an upper bound on $\mathbb{P}(W \notin \mathcal{G})$ using the definition of $\mathcal{G}$ as follows:

$$\mathbb{P}(W \notin \mathcal{G}) \leq \mathbb{P}\left(\|W\|_2 > 2\sqrt{d} + \sqrt{C_1 k \log T}\right)$$
$$+ \mathbb{P}\left(\exists\ 1 \leq i, j \leq N_{\epsilon_0} \text{ s.t. } \left|(x_i^* - x_j^*)^\top W\right| > \sqrt{C_1 k \log T}\, \|x_i^* - x_j^*\|_2\right), \tag{109}$$

leaving us with two terms to control.

- Using the concentration property of χ^2 random variables (e.g., Laurent and Massart (2000, Lemma 1)), we find that the first term on the right-hand side of (109) satisfies

$$\mathbb{P}\left(\|W\|_2 > 2\sqrt{d} + \sqrt{C_1 k \log T}\right) \leq \exp\left\{-\frac{C_1}{2} k \log T\right\}.$$

- When it comes to the second term on the right-hand side of (109), we observe that: for every pair of fixed points x_i^*, x_j^*, one has $\frac{(x_i^* - x_j^*)^\top}{\|x_i^* - x_j^*\|_2} W \sim \mathcal{N}(0,1)$. Thus, it follows from the concentration property of standard Gaussians that

$$\mathbb{P}\left(\left|(x_i^* - x_j^*)^\top W\right| > \sqrt{C_1 k \log T}\, \|x_i^* - x_j^*\|_2\right) = \mathbb{P}\left(\left|\frac{(x_i^* - x_j^*)^\top}{\|x_i^* - x_j^*\|_2} W\right| > \sqrt{C_1 k \log T}\right) \leq \exp\left\{-\frac{C_1}{2} k \log T\right\}.$$

  Combining this with the union-bound and Assumption 1, we can obtain

$$\mathbb{P}\left(\exists\ 1 \leq i, j \leq N_{\epsilon_0} \text{ s.t. } \left|(x_i^* - x_j^*)^\top W\right| > \sqrt{C_1 k \log T}\, \|x_i^* - x_j^*\|_2\right)$$
$$\leq \sum_{1\leq i,j\leq N_{\epsilon_0}} \mathbb{P}\left(\left|(x_i^* - x_j^*)^\top W\right| > \sqrt{C_1 k \log T}\, \|x_i^* - x_j^*\|_2\right) \leq \sum_{1\leq i,j\leq N_{\epsilon_0}} \exp\left\{-\frac{C_1}{2} k \log T\right\}$$
$$\leq N_{\epsilon_0}^2 \exp\left\{-\frac{C_1}{2} k \log T\right\} \leq \exp\left\{\left(2C_{\mathsf{cover}} - C_1/2\right) k \log T\right\} \leq \exp\left\{-\frac{3}{8} C_1 k \log T\right\},$$

  where the last inequality holds provided that $C_1 \geq 16 C_{\mathsf{cover}}$. Consequently, it holds that

$$\mathbb{P}(W \notin \mathcal{G}) \leq \exp\left\{-\frac{C_1}{2} k \log T\right\} + \exp\left\{-\frac{3}{8} C_1 k \log T\right\} \leq 2\exp\left\{-\frac{3}{8} C_1 k \log T\right\}.$$

To finish up, we make the observation that

$$\mathbb{P}(V_\alpha \notin \mathcal{T}_\alpha) \leq \sum_{j\in[N_{\epsilon_0}]\setminus\mathcal{I}} \mathbb{P}(V_1 \in \mathcal{B}_j) + \mathbb{P}(W \notin \mathcal{G})$$
$$\leq 3\exp\left\{-\frac{3}{8} C_1 k \log T\right\} \leq \exp\left\{-\frac{1}{4} C_1 k \log T\right\}.$$

G.2 Proof of Lemma 2

Define the set

$$\mathcal{E}_{\alpha,C}(v) := \left\{ v_1 \,\middle|\, \sqrt{\alpha}\left\|v_1 - x^*_{i(v)}\right\|_2 \geq \sqrt{Ck(1-\alpha)\log T} \right\}.$$

Invoke the Bayes rule to obtain

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_{\alpha,C}(v) \mid V_\alpha = v) &= \frac{\int_{\mathcal{E}_{\alpha,C}(v)} p_{X_0}(v_1) p_{V_\alpha|V_1}(v \mid v_1)\mathrm{d}v_1}{\int p_{X_0}(\widetilde{v}_1) p_{V_\alpha|V_1}(v|\widetilde{v}_1)\mathrm{d}\widetilde{v}_1} \leq \frac{\int_{\mathcal{E}_{\alpha,C}(v)} p_{X_0}(v_1) p_{V_\alpha|V_1}(v \mid v_1)\mathrm{d}v_1}{\int_{\widetilde{v}_1 \in \mathcal{B}_{i(v)}} p_{X_0}(\widetilde{v}_1) p_{V_\alpha|V_1}(v \mid \widetilde{v}_1)\mathrm{d}\widetilde{v}_1} \\
&\leq \frac{\int_{\mathcal{E}_{\alpha,C}(v)} p_{X_0}(v_1) p_{V_\alpha|V_1}(v \mid v_1)\mathrm{d}v_1}{\mathbb{P}(\mathcal{B}_{i(v)}) \inf_{\widetilde{v}_1 \in \mathcal{B}_{i(v)}} p_{V_\alpha|V_1}(v \mid \widetilde{v}_1)} \leq \frac{1}{\mathbb{P}(\mathcal{B}_{i(v)})} \cdot \frac{\sup_{x_0 \in \mathcal{E}_{\alpha,C}(v)} p_{V_\alpha|V_1}(v \mid v_1)}{\inf_{\widetilde{v}_1 \in \mathcal{B}_{i(v)}} p_{V_\alpha|V_1}(v|\widetilde{v}_1)} \\
&\leq \exp(C_1 k \log T) \sup_{v_1 \in \mathcal{E}_{\alpha,C}(v), \widetilde{v}_1 \in \mathcal{B}_{i(v)}} \exp\left\{ \frac{1}{2(1-\alpha)} \left[\left\|v - \sqrt{\alpha}\widetilde{v}_1\right\|_2^2 - \left\|v - \sqrt{\alpha}v_1\right\|_2^2 \right] \right\}.
\end{aligned} \tag{110}$$

Here, the last inequality follows from the property $\mathbb{P}(\mathcal{B}_{i(v)}) \geq \exp\left(-C_1 k \log T\right)$, which is a direct consequence of the assumption $v \in \mathcal{T}_\alpha$.

Further, consider any $(v_1, \widetilde{v}_1)$ with $v_1 \in \mathcal{E}_{\alpha,C}(v)$ and $\widetilde{v}_1 \in \mathcal{B}_{i(v)}$. Without loss of generality, suppose $v_1 \in \mathcal{B}_j$. Using the expression $v = \sqrt{\alpha}v_1^* + \sqrt{1-\alpha}\omega$, we can show that

$$\begin{aligned}
&\left\|v - \sqrt{\alpha}\widetilde{v}_1\right\|_2^2 - \left\|v - \sqrt{\alpha}v_1\right\|_2^2 \\
&\quad = -\alpha \left\|v_1^* - v_1\right\|_2^2 + 2\sqrt{\alpha(1-\alpha)} \left\langle v_1 - \widetilde{v}_1, \omega \right\rangle + \alpha \left\|v_1^* - \widetilde{v}_1\right\|_2^2 \\
&\quad \overset{(a)}{\leq} -\alpha \left(\left\|x^*_{i(x)} - x^*_j\right\|_2 - 2\epsilon_0 \right)^2 + 2\sqrt{\alpha(1-\alpha)} \left\langle v_1 - \widetilde{v}_1, \omega \right\rangle + 4\alpha\epsilon_0^2 \\
&\quad \overset{(b)}{\leq} 4\alpha\epsilon_0 \left\|x^*_{i(v)} - x^*_j\right\|_2 - \alpha \left\|x^*_{i(v)} - x^*_j\right\|_2^2 + 2\sqrt{\alpha(1-\alpha)} \left\{ \left\langle x^*_j - x^*_{i(v)}, \omega \right\rangle + 2(\sqrt{d} + \sqrt{C_1 k \log T})\epsilon_0 \right\}.
\end{aligned}$$

Here, (a) follows from the property of the ϵ_0-net, while (b) combines the definition of the ϵ_0-net with the norm bound for $\omega \in \mathcal{G}$. Moreover, since $\omega \in \mathcal{G}$, it is clearly seen that

$$\left\langle x^*_j - x^*_{i(v)}, \omega \right\rangle \leq \sqrt{C_1 k \log T} \left\|x^*_j - x^*_{i(v)}\right\|_2.$$

In addition, with the choice of ϵ_0 satisfying $\epsilon_0 \ll \sqrt{\frac{1-\alpha}{\alpha}} \min\left\{1, \sqrt{\frac{k \log T}{d}}\right\} \leq \frac{1 \wedge \sqrt{\frac{k \log T}{d}}}{T}$, the following property holds:

$$4\sqrt{\alpha(1-\alpha)}(\sqrt{d} + \sqrt{C_1 k \log T})\epsilon_0 \leq 5(1-\alpha)k \log T.$$

With the preceding bounds in place, we can readily obtain

$$\begin{aligned}
\left\|v - \sqrt{\alpha}\widetilde{v}_1\right\|_2^2 - \left\|v - \sqrt{\alpha}v_1\right\|_2^2 &\leq -\alpha \left\|x^*_{i(v)} - x^*_j\right\|_2^2 + 4(1-\alpha)k \log T \\
&\quad + \left(2\sqrt{C_1 \alpha(1-\alpha)k \log T} + 4\alpha\epsilon_0 \right) \left\|x^*_j - x^*_{i(v)}\right\|_2 \\
&\overset{(a)}{\leq} -\frac{\alpha}{2} \left\|x^*_{i(v)} - x^*_j\right\|_2^2 + 4(1-\alpha)k \log T \\
&\leq -\frac{\alpha}{4} \left\|x^*_{i(v)} - x_0\right\|_2^2 + 4(1-\alpha)k \log T \\
&\leq -\frac{C}{4}(1-\alpha)k \log T + 4(1-\alpha)k \log T \leq -\frac{C}{5}(1-\alpha)k \log T.
\end{aligned}$$

Here, both (a) and the last inequality follow since $v_1 \in \mathcal{E}_{\alpha,C}(v)$ and $C \geq C_2$. Taking the above bound

collectively with (110) yields

$$\mathbb{P}(\mathcal{E}_{\alpha,C}(v) \mid V_\alpha = v) \le \exp(C_1 k \log T) \sup_{x \in \mathcal{E}_{\alpha,C}(v), \widetilde{v}_1 \in \mathcal{B}_{i(v)}} \exp\left\{ \frac{1}{2(1-\alpha)} \left[\left\| v - \sqrt{\alpha}\widetilde{v}_1 \right\|_2^2 - \left\| v - \sqrt{\alpha} v_1 \right\|_2^2 \right] \right\}$$
$$\le \exp(C_1 k \log T) \cdot \exp\left(-\frac{C}{5} \frac{1}{2(1-\alpha)} (1-\alpha) k \log T \right) \le \exp\left(-\frac{C}{20} k \log T \right)$$

as claimed.

G.3 Proof of Lemma 3

To simplify notation of this proof, we shall often employ the following shorthand notation (cf. (42))

$$\mathrm{Cov}_{V_1|V_\alpha} = \mathrm{Cov}_{V_1|V_\alpha}(V_\alpha) \qquad \text{and} \qquad \mathrm{Cov}_{X_0|X_t} = \mathrm{Cov}_{X_0|X_t}(X_t) \tag{111}$$

whenever it is clear from the context.

To bound the term $\mathbb{E}_{X_t}\left[\left\| \mathrm{Cov}_{X_0|X_t} \right\|_{\mathrm{F}}^2 \right]$, we resort to the following result, originally developed in the stochastic localization literature (Eldan, 2020) (see also Benton et al. (2024, Lemma 1)).

Lemma 8 *Let $\lambda_t \coloneqq \sqrt{1-e^{-2t}}$, then for all $t > 0$,*

$$\frac{\lambda_t^3}{2\dot{\lambda}_t} \frac{\mathrm{d}}{\mathrm{d}t} \mathbb{E}_{U_t}\left[\mathrm{Cov}_{U_0|U_t}(U_t) \right] = \mathbb{E}_{U_t}\left[\left(\mathrm{Cov}_{U_0|U_t}(U_t) \right)^2 \right].$$

where $U_t \coloneqq e^{-t} X_0 + \sqrt{1-e^{-2t}} Z$ with $X_0 \sim p_{\mathsf{data}}$ and $Z \sim \mathcal{N}(0, I_d)$. Here, we let $\mathrm{Cov}_{U_1|U_t}(u) = \mathbb{E}[U_1 U_1^\top \mid U_t = u] - \mathbb{E}[U_1 \mid U_t = u]\,\mathbb{E}[U_1 \mid U_t = u]^\top$, and denote by $\dot{\lambda}_t$ the derivative of λ_t with respect to t.

Now, let us introduce the bijection $\alpha(t) \coloneqq e^{-2t}$ that maps $t \in [0, \infty)$ to $\alpha \in (0, 1]$. Define

$$V_\alpha \coloneqq \sqrt{\alpha} X_0 + \sqrt{1-\alpha} Z, \qquad \nu_\alpha \coloneqq \sqrt{1-\alpha}, \qquad \text{and} \qquad t(\alpha) \coloneqq \frac{1}{2} \log \frac{1}{\alpha}. \tag{112}$$

Then it can be readily seen that

$$V_\alpha = U_{t(\alpha)} \qquad \text{and} \qquad \nu_\alpha = \lambda_{t(\alpha)}. \tag{113}$$

Straightforward calculations allow one to rewrite the result in Lemma 8 as

$$\begin{aligned} \mathrm{d}\mathbb{E}_{V_\alpha}\left[\mathrm{Cov}_{V_1|V_\alpha} \right] &= \frac{2}{\lambda_{t(\alpha)}^3} \left. \frac{\mathrm{d}\lambda_t}{\mathrm{d}t} \right|_{t=t(\alpha)} \mathbb{E}_{V_\alpha}\left[\mathrm{Cov}_{V_1|V_\alpha}^2 \right] \mathrm{d}t(\alpha) = \frac{2 \frac{\mathrm{d}\nu_\alpha}{\mathrm{d}\alpha}}{\nu_\alpha^3} \mathbb{E}_{V_\alpha}\left[\mathrm{Cov}_{V_1|V_\alpha}^2 \right] \mathrm{d}\alpha \\ &= -\frac{(1-\alpha)^{-1/2}}{(1-\alpha)^{3/2}} \mathbb{E}_{V_\alpha}\left[\mathrm{Cov}_{V_1|V_\alpha}^2 \right] \mathrm{d}\alpha = -\frac{1}{(1-\alpha)^2} \mathbb{E}_{V_\alpha}\left[\mathrm{Cov}_{V_1|V_\alpha}^2 \right] \mathrm{d}\alpha. \end{aligned} \tag{114}$$

Integrating the above equation over the interval $[\overline{\alpha}_{t+1}, \overline{\alpha}_t)$, we obtain

$$\begin{aligned} \int_{\overline{\alpha}_{t+1}}^{\overline{\alpha}_t} \frac{1}{(1-\alpha)^2} \mathbb{E}_{V_\alpha}\left[\mathrm{Cov}_{V_1|V_\alpha}^2 \right] \mathrm{d}\alpha &= \mathbb{E}_{V_{\overline{\alpha}_{t+1}}}\left[\mathrm{Cov}_{V_1|V_{\overline{\alpha}_{t+1}}} \right] - \mathbb{E}_{V_{\overline{\alpha}_t}}\left[\mathrm{Cov}_{V_1|V_{\overline{\alpha}_t}} \right] \\ &= \mathbb{E}_{X_{t+1}}\left[\mathrm{Cov}_{X_0|X_{t+1}} \right] - \mathbb{E}_{X_t}\left[\mathrm{Cov}_{X_0|X_t} \right]. \end{aligned} \tag{115}$$

Next, we proceed to control the relative magnitude between $\mathrm{Cov}_{V_1|V_\alpha}$ (with $\alpha \in [\overline{\alpha}_{t+1}, \overline{\alpha}_t]$) and $\mathrm{Cov}_{V_1|V_{\overline{\alpha}_{t+1}}}$. To achieve this, we resort to the following SDE to describe the random process $\{\mathrm{Cov}_{V_1|V_\alpha}\}$, whose proof can be found in Eldan (2022, Section 4.2.1):

$$\mathrm{d}\mathrm{Cov}_{V_1|V_\alpha} = -\frac{1}{(1-\alpha)^2} \mathrm{Cov}_{V_1|V_\alpha}^2 \mathrm{d}\alpha + \mathcal{M}_\alpha^{(3)} \mathrm{d}B_{\frac{\alpha}{1-\alpha}}.$$

Here, (B_t) denotes the standard Brownian motion in $\mathbb{R}^d$ and

$$\mathcal{M}_\alpha^{(l)} := \mathbb{E}\left[(V_1 - \mathbb{E}[V_1|V_\alpha])^{\otimes l} \mid V_\alpha\right]. \tag{116}$$

For a vector-valued stochastic process (M_t), we denote its quadratic variation process by $\langle M \rangle$. Apply Itô's formula to obtain

$$\begin{aligned}
\mathrm{d}\left(\mathrm{tr}\left(\mathrm{Cov}^2_{V_1|V_\alpha}\right)\right) &= 2\left\langle \mathrm{Cov}_{V_1|V_\alpha}, \mathrm{dCov}_{V_1|V_\alpha}\right\rangle + \mathrm{d}\left[\mathrm{tr}\left(\left\langle \mathrm{Cov}_{V_1|V_\alpha}\right\rangle\right)\right] \\
&= 2\left\langle \mathrm{Cov}_{V_1|V_\alpha}, \mathcal{M}_\alpha^{(3)} \mathrm{d}B_{\frac{\alpha}{1-\alpha}}\right\rangle - \frac{2}{(1-\alpha)^2}\left\langle \mathrm{Cov}_{V_1|V_\alpha}, \mathrm{Cov}^2_{V_1|V_\alpha}\right\rangle \mathrm{d}\alpha + \frac{1}{(1-\alpha)^2}\langle \mathcal{M}_\alpha^{(3)}, \mathcal{M}_\alpha^{(3)}\rangle \mathrm{d}\alpha.
\end{aligned}$$

Taking expectation then yields

$$\mathrm{d}\left[\mathrm{tr}\left(\mathbb{E}\left[\mathrm{Cov}^2_{V_1|V_\alpha}\right]\right)\right] = -\frac{2}{(1-\alpha)^2}\mathbb{E}\left[\left\langle \mathrm{Cov}_{V_1|V_\alpha}, \mathrm{Cov}^2_{V_1|V_\alpha}\right\rangle\right]\mathrm{d}\alpha + \frac{1}{(1-\alpha)^2}\mathbb{E}\left[\langle \mathcal{M}_\alpha^{(3)}, \mathcal{M}_\alpha^{(3)}\rangle\right]\mathrm{d}\alpha. \tag{117}$$

We now analyze $\mathbb{E}\left[\left\langle \mathrm{Cov}_{V_1|V_\alpha}, \mathrm{Cov}^2_{V_1|V_\alpha}\right\rangle\right]$. Towards this, use the symmetry of $\mathrm{Cov}_{V_1|V_\alpha}$ to derive that

$$\begin{aligned}
\mathbb{E}\left[\left\langle \mathrm{Cov}_{V_1|V_\alpha}, \mathrm{Cov}^2_{V_1|V_\alpha}\right\rangle\right] &= \mathbb{E}\left[\mathrm{tr}\left(\mathrm{Cov}^3_{V_1|V_\alpha}\right)\right] \le \mathbb{E}\left[\left\|\mathrm{Cov}_{V_1|V_\alpha}\right\| \cdot \left\|\mathrm{Cov}_{V_1|V_\alpha}\right\|_{\mathrm{F}}^2\right] \\
&\le \mathbb{E}\left[\mathrm{tr}(\mathrm{Cov}_{V_1|V_\alpha}) \cdot \left\|\mathrm{Cov}_{V_1|V_\alpha}\right\|_{\mathrm{F}}^2\right] \\
&= \mathbb{E}\left[\mathrm{tr}(\mathrm{Cov}_{V_1|V_\alpha})\mathbb{1}\{V_\alpha \in \mathcal{T}_\alpha\} \cdot \left\|\mathrm{Cov}_{V_1|V_\alpha}\right\|_{\mathrm{F}}^2\right] + \mathbb{E}\left[\mathrm{tr}(\mathrm{Cov}_{V_1|V_\alpha})\mathbb{1}\{V_\alpha \notin \mathcal{T}_\alpha\} \cdot \left\|\mathrm{Cov}_{V_1|V_\alpha}\right\|_{\mathrm{F}}^2\right] \\
&\le C_3\frac{1-\alpha}{\alpha}(k\log T)\mathbb{E}\left[\left\|\mathrm{Cov}_{V_1|V_\alpha}\right\|_{\mathrm{F}}^2\right] + \mathbb{E}\left[\mathrm{tr}(\mathrm{Cov}_{V_1|V_\alpha})\mathbb{1}\{V_\alpha \notin \mathcal{T}_\alpha\} \cdot \left\|\mathrm{Cov}_{V_1|V_\alpha}\right\|_{\mathrm{F}}^2\right].
\end{aligned}$$

Here, the last inequality follows from the definition of $\mathcal{T}_\alpha$ in (41) and Corollary 1. Regarding the last term of the above inequality, combining Assumption 2 and Lemma 1 results in

$$\mathbb{E}\left[\mathrm{tr}(\mathrm{Cov}_{V_1|V_\alpha})\mathbb{1}\{V_\alpha \notin \mathcal{T}_\alpha\} \cdot \left\|\mathrm{Cov}_{V_1|V_\alpha}\right\|_{\mathrm{F}}^2\right] \le 8T^{6c_R}\mathbb{P}(V_\alpha \notin \mathcal{T}_\alpha) \le \frac{1}{T^{10}}.$$

As a consequence, we arrive at

$$\mathbb{E}\left[\left\langle \mathrm{Cov}_{V_1|V_\alpha}, \mathrm{Cov}^2_{V_1|V_\alpha}\right\rangle\right] \le C_3\frac{1-\alpha}{\alpha}(k\log T)\mathbb{E}\left[\left\|\mathrm{Cov}_{V_1|V_\alpha}\right\|_{\mathrm{F}}^2\right] + \frac{1}{T^{10}}. \tag{118}$$

Substitution into (117) yields

$$\begin{aligned}
\mathrm{d}\mathbb{E}\left[\left\|\mathrm{Cov}_{V_1|V_\alpha}\right\|_{\mathrm{F}}^2\right] &= -\frac{2}{(1-\alpha)^2}\mathbb{E}\left[\left\langle \mathrm{Cov}_{V_1|V_\alpha}, \mathrm{Cov}^2_{V_1|V_\alpha}\right\rangle\right]\mathrm{d}\alpha + \frac{1}{(1-\alpha)^2}\mathbb{E}\left[\left\langle \mathcal{M}_\alpha^{(3)}, \mathcal{M}_\alpha^{(3)}\right\rangle\right]\mathrm{d}\alpha \\
&\ge -\frac{2}{(1-\alpha)^2}\mathbb{E}\left[\left\langle \mathrm{Cov}_{V_1|V_\alpha}, \mathrm{Cov}^2_{V_1|V_\alpha}\right\rangle\right]\mathrm{d}\alpha \\
&\ge -\frac{2}{(1-\alpha)^2}C_3\frac{1-\alpha}{\alpha}(k\log T)\mathbb{E}\left[\left\|\mathrm{Cov}_{V_1|V_\alpha}\right\|_{\mathrm{F}}^2\right]\mathrm{d}\alpha - \frac{1}{T^{10}}\mathrm{d}\alpha \\
&\ge -\frac{2C_3 k\log T}{\overline{\alpha}_{t+1}(1-\overline{\alpha}_t)}\mathbb{E}\left[\left\|\mathrm{Cov}_{V_1|V_\alpha}\right\|_{\mathrm{F}}^2\right]\mathrm{d}\alpha - \frac{1}{T^{10}}\mathrm{d}\alpha,
\end{aligned}$$

where the last line holds since $\alpha \in [\overline{\alpha}_{t+1}, \overline{\alpha}_t]$. In view of Grownwall's inequality, we can derive

$$\begin{aligned}
\exp\left\{\frac{2C_3 k\alpha\log T}{\overline{\alpha}_{t+1}(1-\overline{\alpha}_t)}\right\}\mathbb{E}\left[\left\|\mathrm{Cov}_{V_1|V_\alpha}\right\|_{\mathrm{F}}^2\right] - \exp\left\{\frac{2C_3 k\overline{\alpha}_{t+1}\log T}{\overline{\alpha}_{t+1}(1-\overline{\alpha}_t)}\right\}\mathbb{E}\left[\left\|\mathrm{Cov}_{V_1|V_{\overline{\alpha}_{t+1}}}\right\|_{\mathrm{F}}^2\right] \\
\ge -\frac{1}{T^{10}}\int_{\overline{\alpha}_{t+1}}^{\alpha}\exp\left\{\frac{2C_3 k\alpha'\log T}{\overline{\alpha}_{t+1}(1-\overline{\alpha}_t)}\right\}\mathrm{d}\alpha'
\end{aligned}$$

$$= -\frac{\overline{\alpha}_{t+1}(1-\overline{\alpha}_t)}{2C_3kT^{10}\log T}\left(\exp\left\{\frac{2C_3k\alpha\log T}{\overline{\alpha}_{t+1}(1-\overline{\alpha}_t)}\right\} - \exp\left\{\frac{2C_3k\overline{\alpha}_{t+1}\log T}{\overline{\alpha}_{t+1}(1-\overline{\alpha}_t)}\right\}\right).$$

Dividing both sides of the above inequality by $\exp\left\{\frac{2C_3k\overline{\alpha}_{t+1}\log T}{\overline{\alpha}_{t+1}(1-\overline{\alpha}_t)}\right\}$, we obtain

$$\begin{aligned}\exp\left\{\frac{2C_3k(\alpha-\overline{\alpha}_{t+1})\log T}{\overline{\alpha}_{t+1}(1-\overline{\alpha}_t)}\right\}&\mathbb{E}\left[\left\|\mathrm{Cov}_{V_1|V_\alpha}\right\|_{\mathrm{F}}^2\right] - \mathbb{E}\left[\left\|\mathrm{Cov}_{V_1|V_{\overline{\alpha}_t}}\right\|_{\mathrm{F}}^2\right]\\ &\geq -\frac{\overline{\alpha}_{t+1}(1-\overline{\alpha}_t)}{2C_3kT^{10}\log T}\left(\exp\left\{\frac{2C_3k(\alpha-\overline{\alpha}_{t+1})\log T}{\overline{\alpha}_{t+1}(1-\overline{\alpha}_t)}\right\} - 1\right).\end{aligned}$$

According to Lemma 4, any $\alpha \in [\overline{\alpha}_{t+1}, \overline{\alpha}_t]$ obeys

$$\frac{2C_3k(\alpha-\overline{\alpha}_{t+1})\log T}{\overline{\alpha}_{t+1}(1-\overline{\alpha}_t)} \leq \frac{2C_3k(1-\alpha_{t+1})\log T}{\alpha_{t+1}-\overline{\alpha}_{t+1}} \leq \frac{8C_3c_1k\log^2 T}{T} \leq 1,$$

provided that $8C_3c_1k\log^2 T \leq T$. Consequently, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\mathrm{Cov}_{V_1|V_{\overline{\alpha}_{t+1}}}\right\|_{\mathrm{F}}^2\right] &\leq \exp\left\{\frac{2C_3k(\alpha-\overline{\alpha}_{t+1})\log T}{\overline{\alpha}_{t+1}(1-\overline{\alpha}_t)}\right\}\mathbb{E}\left[\left\|\mathrm{Cov}_{V_1|V_\alpha}\right\|_{\mathrm{F}}^2\right]\\ &\quad + \frac{\overline{\alpha}_{t+1}(1-\overline{\alpha}_t)}{2C_3kT^{10}\log T}\left(\exp\left\{\frac{2C_3k(\alpha-\overline{\alpha}_{t+1})\log T}{\overline{\alpha}_{t+1}(1-\overline{\alpha}_t)}\right\} - 1\right)\\ &\leq 3\mathbb{E}\left[\left\|\mathrm{Cov}_{V_1|V_\alpha}\right\|_{\mathrm{F}}^2\right] + \frac{\overline{\alpha}_{t+1}(1-\overline{\alpha}_t)}{2C_3kT^{10}\log T}\left(\exp\left\{\frac{2C_3k(\alpha-\overline{\alpha}_{t+1})\log T}{\overline{\alpha}_{t+1}(1-\overline{\alpha}_t)}\right\} - 1\right)\\ &\overset{(a)}{\leq} 3\mathbb{E}\left[\left\|\mathrm{Cov}_{V_1|V_\alpha}\right\|_{\mathrm{F}}^2\right] + \frac{2(\overline{\alpha}_t-\overline{\alpha}_{t+1})}{T^{10}},\end{aligned} \tag{119}$$

where (a) holds since $e^x - 1 \leq 2x$ for all $x \leq 1$. Combining (115) and (119), and making use of the equivalence between X_t and $V_{\overline{\alpha}_t}$, we arrive at

$$\begin{aligned}\mathbb{E}\left[\mathrm{tr}\left(\mathrm{Cov}_{X_0|X_{t+1}}\right)\right] - &\mathbb{E}\left[\mathrm{tr}\left(\mathrm{Cov}_{X_0|X_t}\right)\right] = \int_{\overline{\alpha}_{t+1}}^{\overline{\alpha}_t}\frac{1}{(1-\alpha)^2}\mathbb{E}\left[\mathrm{tr}\left(\mathrm{Cov}^2_{V_1|V_\alpha}\right)\right]\\ &\geq \int_{\overline{\alpha}_{t+1}}^{\overline{\alpha}_t}\frac{1}{3(1-\alpha)^2}\left\{\mathbb{E}\left[\left\|\mathrm{Cov}^2_{V_1|V_{\overline{\alpha}_{t+1}}}\right\|_{\mathrm{F}}^2\right] - \frac{2(\overline{\alpha}_t-\overline{\alpha}_{t+1})}{T^{10}}\right\}\mathrm{d}\alpha\\ &= \frac{\overline{\alpha}_t(1-\alpha_{t+1})}{3(1-\overline{\alpha}_t)(1-\overline{\alpha}_{t+1})}\mathbb{E}\left[\left\|\mathrm{Cov}_{X_0|X_{t+1}}\right\|_{\mathrm{F}}^2\right] - \frac{\overline{\alpha}_t^2(1-\alpha_{t+1})^2}{T^{10}(1-\overline{\alpha}_t)(1-\overline{\alpha}_{t+1})}.\end{aligned}$$

This in turn allows us to derive

$$\begin{aligned}\widetilde{\sigma}_t^2\mathbb{E}_{X_t}\left[\left\|\mathrm{Cov}_{X_0|X_t}\right\|_{\mathrm{F}}^2\right] &= \frac{(1-\alpha_t)\overline{\alpha}_t}{(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)}\mathbb{E}_{X_t}\left[\mathrm{tr}\left(\mathrm{Cov}^2_{X_0|X_t}\right)\right]\\ &\leq 3\left\{\mathbb{E}\left[\mathrm{tr}(\mathrm{Cov}_{X_0|X_t})\right] - \mathbb{E}\left[\mathrm{tr}(\mathrm{Cov}_{X_0|X_{t-1}})\right]\right\} + \frac{3}{T^{10}},\end{aligned} \tag{120}$$

thus completing the proof of this lemma.

G.4 Proof of Lemma 4

A little algebra yields

$$\begin{aligned}\frac{\overline{\alpha}_t(1-\alpha_t)}{2(\alpha_t-\overline{\alpha}_t)(1-\overline{\alpha}_t)} - \frac{\overline{\alpha}_{t+1}(1-\alpha_{t+1})}{2(\alpha_{t+1}-\overline{\alpha}_{t+1})(1-\overline{\alpha}_{t+1})} &= \frac{\overline{\alpha}_{t-1}(1-\alpha_t)(1-\overline{\alpha}_{t+1}) - \overline{\alpha}_t(1-\alpha_{t+1})(1-\overline{\alpha}_{t-1})}{2(1-\overline{\alpha}_{t-1})(1-\overline{\alpha}_t)(1-\overline{\alpha}_{t+1})}\\ &\overset{(a)}{\leq} \frac{\overline{\alpha}_{t-1}(1-\alpha_t)[(1-\overline{\alpha}_{t+1}) - \alpha_t + \overline{\alpha}_t]}{2(1-\overline{\alpha}_{t-1})(1-\overline{\alpha}_t)(1-\overline{\alpha}_{t+1})} = \frac{\overline{\alpha}_{t-1}(1-\alpha_t)[1-\alpha_t+\overline{\alpha}_t(1-\alpha_{t+1})]}{2(1-\overline{\alpha}_{t-1})(1-\overline{\alpha}_t)(1-\overline{\alpha}_{t+1})}\\ &\leq \frac{\overline{\alpha}_{t-1}(1-\alpha_t)(1-\alpha_{t+1})}{(1-\overline{\alpha}_{t-1})(1-\overline{\alpha}_t)(1-\overline{\alpha}_{t+1})} \leq \left(\frac{8c_1\log T}{T}\right)^2\frac{\overline{\alpha}_t}{1-\overline{\alpha}_t},\end{aligned}$$

where (a) follows since $1-\alpha_t \le 1-\alpha_{t+1}$, and the last inequality applies (46).

G.5 Proof of Lemma 5

Note that for any matrix B, we know that B and $B^\top$ have the same determinant. As a result,

$$\begin{aligned}\log\det(I+\eta A+\eta\Delta) &= \frac{1}{2}\left\{\log\det(I+\eta A+\eta\Delta^\top)+\log\det(I+\eta A+\eta\Delta)\right\}\\ &= \frac{1}{2}\log\det\left(I+2\eta A+\eta(\Delta^\top+\Delta)+\eta^2(A+\Delta)^\top(A+\Delta)\right).\end{aligned} \tag{121}$$

For any vector $x\in\mathbb{R}^d$, we observe that

$$\begin{aligned}x^\top\big(I+\eta(2A+\Delta^\top+\Delta)\big)x &= \|x\|_2^2+2\eta x^\top Ax+\eta x^\top(\Delta^\top+\Delta)x\\ &\ge \|x\|_2^2-\eta\left\|\Delta^\top+\Delta\right\|\|x\|_2^2 \ge (1-2\eta\|\Delta\|)\|x\|_2^2 \ge \frac{1}{2}\|x\|_2^2,\end{aligned}$$

which implies that the matrix $I+2\eta A+\eta(\Delta^\top+\Delta)\succ 0$. Further, it is easily seen that

$$I+2\eta A+\eta(\Delta^\top+\Delta)\preceq I+2\eta A+\eta(\Delta^\top+\Delta)+\eta^2(A+\Delta)^\top(A+\Delta).$$

According to the Löwner–Heinz theorem, $\log A\preceq\log B$ holds for any $0\preceq A\preceq B$. This in turn allows one to derive

$$\begin{aligned}&\log\det\left(I+2\eta A+\eta(\Delta^\top+\Delta)+\eta^2(A+\Delta)^\top(A+\Delta)\right)\\ &\qquad= \mathrm{tr}\Big(\log\left(I+2\eta A+\eta(\Delta^\top+\Delta)+\eta^2(A+\Delta)^\top(A+\Delta)\right)\Big)\\ &\qquad\ge \mathrm{tr}\Big(\log\left(I+2\eta A+\eta(\Delta^\top+\Delta)\right)\Big)=\log\det\left(I+2\eta A+\eta(\Delta^\top+\Delta)\right).\end{aligned} \tag{122}$$

For any symmetric matrix $B\in\mathbb{R}^{d\times d}$, we denote its eigenvalues as $\{\lambda_i(B)\}_{i=1}^d$. Then according to Weyl's inequality, we have

$$\lambda_i\big(2\eta A+\eta(\Delta^\top+\Delta)\big)\ge 2\eta\lambda_i(A)-\eta\left\|\Delta^\top+\Delta\right\|\ge -2\eta\|\Delta\|\ge -\frac{1}{2}\qquad\text{for all } i\le d.$$

Further, it can be verified that $\log(1+x)\ge x-x^2$ holds for any $x\ge -1/2$, which results in

$$\begin{aligned}\log\det\left(I+2\eta A+\eta(\Delta^\top+\Delta)\right) &\ge \sum_{i=1}^d \eta\lambda_i\left(2A+\Delta^\top+\Delta\right)-\sum_{i=1}^d\eta^2\lambda_i^2\left(2A+\Delta^\top+\Delta\right)\\ &= \eta\mathrm{tr}\left(2A+\Delta^\top+\Delta\right)-\eta^2\left\|2A+\Delta^\top+\Delta\right\|_{\mathrm{F}}^2\\ &\ge 2\eta\mathrm{tr}(A)+2\eta\mathrm{tr}(\Delta)-8\eta^2\|A\|_{\mathrm{F}}^2-8\eta^2\|\Delta\|_{\mathrm{F}}^2.\end{aligned} \tag{123}$$

The proof can thus be completed by combining (121), (122) and (123).