

Convergence of the denoising diffusion probabilistic models under general noise schedules

Yumiharu Nakano^{*1}

¹Department of Mathematical and Computing Science
Institute of Science Tokyo

August 5, 2025

Abstract

This paper presents a theoretical convergence analysis of a denoising diffusion probabilistic model (DDPM) in its original discrete-time formulation introduced by Ho, Jain, and Abbeel (*Advances in Neural Information Processing Systems*, **33** (2020), 6840–6851). We derive an explicit upper bound for the total variation distance between the sampling distribution of the discrete-time DDPM algorithm and a given target data distribution, under general noise schedule parameters. Our analysis requires only mild regularity assumptions on the data distribution and a linear growth condition on the estimated score function. The sampling scheme is interpreted as an exponential-integrator-type approximation of a reverse-time stochastic differential equation (SDE) over a finite time horizon. Tools from the Schrödinger problem are employed to control the distributional error in reverse time and connect it to its forward-time counterpart. Moreover, the score function in DDPMs naturally appears as an adapted solution of a forward-backward SDE, providing a basis for analyzing the time-discretization error in reverse-time SDE sampling.

Key words: Denoising diffusion probabilistic model, reverse-time stochastic differential equations, Schrödinger problem, forward-backward stochastic differential equations.

AMS MSC 2020: 60H30, 65C30, 60J60

1 Introduction

Denoising diffusion probabilistic models (DDPMs), introduced by Ho, Jain and Abbeel [10], are a class of diffusion-based generative models, initiated by Sohl-Dickstein et al. [34], that have achieved remarkable empirical success in diverse application areas such as computer vision [11, 21, 25, 28, 30, 31, 33, 41, 43], medical imaging [5, 29, 35], time-series generation [37, 24], audio and speech synthesis [2, 14, 12, 23], and computational chemistry [17, 26, 39]; see also [1, 40] for surveys.

A DDPM consists of two stages: (i) a forward Markov process that gradually perturbs data into a tractable noise distribution (typically a Gaussian), and (ii) a reverse-time dynamics that

^{*}E-mail: nakano@comp.isct.ac.jp

transforms pure noise back into data samples. In continuous-time formulations [36], these dynamics are described by stochastic differential equations (SDEs), leading to efficient samplers based on probability flow ODEs and exponential integrators [42]. A crucial part of the reverse dynamics is the learned score function that predicts the injected noise.

From a probabilistic perspective, the reverse-time DDPM dynamics can be viewed as an approximation of a reverse-time SDE defined over a finite horizon. Such SDEs are closely connected to the Schrödinger problem [18, 4], which seeks the most likely stochastic evolution between given end-point distributions, and to forward-backward SDEs (FBSDEs), which naturally appear in stochastic control. This connection highlights DDPMs as an applied framework where classical probabilistic objects (Schrödinger bridges, FBSDEs, time reversal of diffusions) arise naturally, but with additional difficulties such as discrete-time approximation and learned, imperfect scores. Analyzing DDPMs therefore provides a unique opportunity to bridge modern machine learning practice and established probabilistic theory.

In this paper, we provide a theoretical analysis of the original discrete-time DDPM sampling algorithm under *general noise schedules*, without structural restrictions on their functional form. We derive an explicit upper bound on the total variation distance between the sampling distribution and the target data distribution, under mild regularity conditions on the data and a linear growth condition on the score estimate. Our analysis reveals that the sampling error can be decomposed into three parts: the error induced by the data distribution itself relative to the Gaussian reference law, the score-matching error due to imperfect training, and the time discretization error associated with approximating the reverse-time SDE via an exponential-integrator-type scheme.

Related work

Several recent works investigate diffusion-type generative models from a mathematical viewpoint. On the continuous-time side, De Bortoli et al. [7] and De Bortoli [6] derived total variation error bounds for exponential-integrator-type discretizations of reverse-time SDEs under mild regularity assumptions, but their results involve restrictive relationships between time-step size, score error, and terminal time. Lee et al. [15] proved convergence of an Euler–Maruyama scheme for reverse-time SDEs assuming that the target density satisfies a log-Sobolev inequality, which essentially excludes multimodal distributions. A subsequent work [16] studied algorithms similar to discrete-time DDPMs but required nonstandard initialization and additional cutoffs. Other analyses such as [3] considered specific (constant) noise schedules in continuous time, while [19, 20] investigated discrete-time variants but derived bounds only for intermediate-time states rather than the final output. Mbacke and Rivasplata [27] obtained 1-Wasserstein error bounds for discrete-time DDPMs but adopted a learning objective different from the usual score-matching approach.

In contrast, the present work establishes, for the first time to our knowledge, a total variation convergence bound for the original discrete-time DDPM algorithm under *general noise schedules*, making essential use of Schrödinger bridge techniques for reverse-time error control and an FBSDE representation of the score function.

Contributions

The main contributions of this paper are:

- A probabilistic interpretation of the discrete-time DDPM sampling procedure as an exponential-integrator-type approximation of a reverse-time SDE, linked to the Schrödinger problem and FBSDE theory.
- An explicit total variation error bound separating contributions from score-matching error and time discretization.
- A discussion of practically used noise schedules, showing that our mild assumptions are satisfied in common implementations.

Organization of the paper

Section 2 presents the main convergence result and its proof sketch. Section 3 contains technical lemmas and detailed proofs.

2 Main results

2.1 Notation

Denote by ∇ the gradient operator. We often write ∇_x for the gradient with respect to the variable x . For a function f on $[0, 1] \times \mathbb{R}^d$ we denote by ∇f the gradient of f with respect to the spatial variable. We denote by $\partial_t f$ and $\partial_{x_j} f$ the partial derivatives of $f(t, x)$ with respect to the time variable t and j -th component x_j of the spatial variable x respectively. Let $\mathcal{P}(\mathcal{X})$ be the set of all Borel probability measures on a Polish space \mathcal{X} . Denote by a^\top the transpose of a vector or matrix a .

2.2 Results

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space. Let $\mu_{data} \in \mathcal{P}(\mathbb{R}^d)$ and $\{\alpha_i\}_{i=1}^n$ be a sequence such that $\alpha_i \in (0, 1)$, $i = 1, \dots, n$. Let \mathbf{x}_0 and Z be random variables with $\mathbf{x}_0 \sim \mu_{data}$ and $Z \sim N(0, I_d)$. The forward Markovian dynamics $\{\mathbf{x}_i\}_{i=0}^n$ is described by

$$\mathbf{x}_i = \sqrt{\alpha_i} \mathbf{x}_{i-1} + \sqrt{1 - \alpha_i} Z_i, \quad i = 1, \dots, n,$$

where $\{Z_i\}_{i=1}^n$ is an IID sequence with $Z_1 \sim N(0, I_d)$ that is independent of \mathbf{x}_0 . In other words, the conditional density $\mathbf{p}_i(x | \mathbf{x}_{i-1})$ of \mathbf{x}_i given \mathbf{x}_{i-1} is the Gaussian density function of x with mean vector $\sqrt{\alpha_i} \mathbf{x}_{i-1}$ and variance-covariance matrix $(1 - \alpha_i) I_d$, $i = 1, \dots, n$. Then

$$\mathbf{x}_i \sim \sqrt{\alpha_i} \mathbf{x}_0 + \sqrt{1 - \alpha_i} Z$$

for each $i = 1, \dots, n$.

Let $\{z_i\}_{i=1}^n$ be a sequence of Borel measurable functions on \mathbb{R}^d , which is interpreted as the resulting denoising term in DDPM algorithm. Let $\{\xi_i\}_{i=1}^n$ be an IID sequence on $(\Omega, \mathcal{F}, \mathbb{P})$ with common distribution $N(0, I_d)$. Define the sequence $\{\mathbf{x}_i^*\}_{i=0}^n$ of random variables by

$$(1) \quad \begin{cases} \mathbf{x}_n^* = \xi_n, \\ \mathbf{x}_{i-1}^* = \frac{1}{\sqrt{\alpha_i}} \left(\mathbf{x}_i^* - \frac{1 - \alpha_i}{\sqrt{1 - \alpha_i}} z_i(\mathbf{x}_i^*) \right) + \sigma_i \xi_i, \quad i \in \{1, \dots, n\}. \end{cases}$$

where $\bar{\alpha}_i = \prod_{k=1}^i \alpha_k$ and $\sigma_i^2 = (1 - \alpha_i)/\alpha_i$.

Remark. In the original DDPM algorithm [10], no additional noise is injected at the final sampling step, while in some variants one more Gaussian perturbation is added. From a probabilistic viewpoint, this difference affects only the last iteration and thus induces an error of the same order as one time step of the discretization. Since our convergence bound already accounts for the overall time discretization error, we do not distinguish between these two variants in the analysis.

The learning objective in (2) is formulated in this framework as follows:

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E} |Z - z_i(\sqrt{\bar{\alpha}_i} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_i} Z)|^2,$$

which is the simplified version of the objective derived from the variational lower bound of the negative of log likelihood of generative models (see [10]). It is also known that this objective is equivalent to the score-matching one. More precisely, with the function

$$\mathbf{s}_i(x) := -\frac{1}{\sqrt{1 - \bar{\alpha}_i}} z_i(x)$$

we get

$$(2) \quad \mathbb{E} |\mathbf{s}_i(\mathbf{x}_i) - \nabla \log \mathbf{p}_i(\mathbf{x}_i)|^2 = \frac{1}{1 - \bar{\alpha}_i} \mathbb{E} |z_i(\mathbf{x}_i) - Z|^2 + \mathbb{E} [|\nabla \log \mathbf{p}_i(\mathbf{x}_i | \mathbf{x}_0)|^2 - |\nabla \log \mathbf{p}_i(\mathbf{x}_i)|^2],$$

where \mathbf{p}_i is the density of \mathbf{x}_i and the score function $\nabla \log \mathbf{p}_i(\cdot)$ of \mathbf{x}_i , $i = 1, \dots, n$, is defined by

$$(3) \quad \nabla \log \mathbf{p}_i(x) = \begin{cases} \nabla \mathbf{p}_i(x) / \mathbf{p}_i(x), & \text{if } \mathbf{p}_i(x) > 0, \\ 0 & \text{otherwise} \end{cases}$$

(see [3] and Section 2.3 below for a proof). Then the score-matching error L in Section 1 is represented as

$$L = \frac{1}{n} \sum_{i=1}^n \mathbb{E} |\mathbf{s}_i(\mathbf{x}_i) - \nabla \log \mathbf{p}_i(\mathbf{x}_i)|^2.$$

We make the following condition on μ_{data} :

(H1) μ_{data} has a bounded density p_{data} such that ∇p_{data} exists in the distribution sense and

$$|\nabla p_{data}(x) + p_{data}(x) Q x| \leq c_0 p_{data}(x), \quad \text{a.e. } x \in \mathbb{R}^d$$

for some $c_0 > 0$ and some symmetric positive definite matrix $Q \in \mathbb{R}^{d \times d}$.

Remark. Suppose that

$$p_{data}(x) \propto e^{-\frac{1}{2} x^\top Q x - U(x)}, \quad x \in \mathbb{R}^d,$$

where U is Lipschitz continuous on \mathbb{R}^d with Lipschitz constant c_0 . Then, by the well-known Rademacher's theorem (see, e.g., Evans [8]), we have $|\nabla U| \leq c_0$ a.e., and so

$$|\nabla p_{data}(x) + p_{data}(x) Q x| \leq |\nabla U(x) p_{data}(x)| \leq c_0 p_{data}(x), \quad \text{a.e. } x \in \mathbb{R}^d.$$

Thus in this case (H1) is satisfied. This is also true for the density p_{data} of the form

$$p_{data}(x) \propto e^{-\frac{1}{2}x^\top Qx - U(x)} 1_S(x), \quad x \in \mathbb{R}^d,$$

where S is a bounded open subset of \mathbb{R}^d with Lipschitz boundary.

Let c_1 be a given positive constant that is greater than the maximum eigenvalue of Q . The condition (H1) leads to a linear growth of the score function $\nabla \log \mathbf{p}_i(x)$.

Lemma 1. *The function $\nabla \log \mathbf{p}_i(x)$ satisfies*

$$|\nabla \log \mathbf{p}_i(x)| \leq \frac{c_0}{\sqrt{\bar{\alpha}_i}} + \frac{c_1}{\bar{\alpha}_i} |x|, \quad x \in \mathbb{R}^d, \quad i = 1, \dots, n.$$

The condition (H1) and Lemma 1 suggest that it is natural to assume that the estimated score function \mathbf{s}_i satisfies the same growth condition as that for $\nabla \log \mathbf{p}_i$.

(H2) The function \mathbf{s}_i or z_i , $i = 0, 1, \dots, n-1$, satisfies

$$|\mathbf{s}_i(x)| \leq \frac{c_0}{\sqrt{\bar{\alpha}_i}} + \frac{c_1}{\bar{\alpha}_i} |x|, \quad \text{a.e. } x \in \mathbb{R}^d, \quad i = 1, \dots, n.$$

Remark. One might be concerned that (H2) is not automatically satisfied, as the function \mathbf{s}_i is typically defined by a neural network. However, since c_0 and c_1 can be taken arbitrarily large, the condition (H2) is practically harmless. Theoretically, it is possible to redefine the function \mathbf{s}_i in such a way that (H2) is fulfilled without increasing the learning error. Indeed, consider

$$\tilde{\mathbf{s}}_i(x) := \mathbf{s}_i(x) 1_{\{|\mathbf{s}_i(x)| \leq B_i(x)\}} + \nabla \log \mathbf{p}_i(x) 1_{\{|\mathbf{s}_i(x)| > B_i(x)\}}, \quad x \in \mathbb{R}^d,$$

where $B_i(x) = c_0/\sqrt{\bar{\alpha}_i} + (c_1/\bar{\alpha}_i)|x|$. It follows from Lemma 1 that $|\tilde{\mathbf{s}}_i(x)| \leq B_i(x)$ and $|\tilde{\mathbf{s}}_i(x) - \nabla \log \mathbf{p}_i(x)| \leq |\mathbf{s}_i(x) - \nabla \log \mathbf{p}_i(x)|$.

To estimate the other weak approximation errors, we adopt the total variation distance D_{TV} defined by

$$D_{TV}(\mu, \nu) = \sup_{\|f\|_\infty \leq 1} \left| \int_{\mathbb{R}^d} f(x) (\mu - \nu)(dx) \right|, \quad \mu, \nu \in \mathcal{P}(\mathbb{R}^d),$$

where $\|f\|_\infty = \sup_{x \in \mathbb{R}^d} |f(x)|$.

Put $\alpha_{min} = \min_{1 \leq i \leq n} \alpha_i$. Here is the main result of this paper.

Theorem 1. *Suppose that (H1) and (H2) hold. Then there exist a constant $C > 0$, only depending on c_0 , c_1 , and $\mathbb{E}|\mathbf{x}_0|^2$, and a constant $\delta > 0$ such that if $\bar{\alpha}_n < \delta$ then*

$$(4) \quad \begin{aligned} & D_{TV}(\mu_{data}, \mathbb{P}(\mathbf{x}_0^*)^{-1}) \\ & \leq C \sqrt{d\sqrt{\bar{\alpha}_n} + \sqrt{d}(-n \log \alpha_{min})(\bar{\alpha}_n)^{-2}\sqrt{L} + d^2 e^{c_2(\bar{\alpha}_n)^{-1}} n (\log \alpha_{min})^2}, \end{aligned}$$

where $c_2 = 10c_0 + 7c_1 + 1$.

As we see below, the 1st term of the right-hand side in (4) comes from the Langevin error, i.e., the distributional difference between \mathbf{x}_n and a standard Gaussian random variable. The 2nd term comes from the score estimation process, and the 3rd term is the time discretization error for the reverse-time SDE.

There exists a trade-off with respect to $\bar{\alpha}_n$ on the right-hand sides of Theorem 1. To clarify the convergence of the distributional distance, we describe the behavior of the decreasing noise schedule parameter α_i over time steps and simplify the right-hand side expressions.

Corollary 1. *Suppose that (H1) and (H2) hold. Suppose moreover that*

$$(5) \quad \frac{\gamma_1 \log \log \log n}{n} \leq -\log \alpha_i \leq \frac{\gamma_2 \log \log \log n}{n}, \quad i = 1, \dots, n,$$

for some $\gamma_1, \gamma_2 > 0$. Then for any $\varepsilon > 0$ there exists $n_0 \in \mathbb{N}$ such that for any $n \geq n_0$

$$\begin{aligned} D_{TV}(\mu_{data}, \mathbb{P}(\mathbf{x}_0^*)^{-1}) \\ \leq C \sqrt{d(\log \log n)^{-\gamma_1/2} + \sqrt{d}\gamma_2(\log \log n)^{2\gamma_2+1}\sqrt{L} + d^2\gamma_2 n^{-(1-\varepsilon)}(\log \log \log n)^2} \end{aligned}$$

for some constant $C > 0$, only depending on c_0, c_1 , and $\mathbb{E}|\mathbf{x}_0|^2$.

Remark. In [10], the case of $n = 1000$ is examined and the variances $1 - \alpha_i$ of the forward process are set to be increasing linearly from $1 - \alpha_1 = 10^{-4}$ to $1 - \alpha_n = 0.02$. Thus (5) holds with $\gamma_1 = 0.15$ and $\gamma_2 = 30.67$. Given that these constants have plausible values, the condition (5) aligns with the practical noise schedules used in DDPMs.

2.3 Proof sketch

Here we outline a proof of Theorem 1. Throughout this section we assume (H1) and (H2). Denote by C generic constants only depending on c_0, c_1 , and $\mathbb{E}|\mathbf{x}_0|^2$, which may vary from line to line. First, let us represent $\hat{\mathbf{x}}$ as an exponential integrator type time discretization of a reverse-time SDE. To this end, take the linear interpolation $g(t)$ of $\{0, -\log \alpha_1, \dots, -\sum_{i=1}^n \log \alpha_i\}$ on $\{t_0, t_1, \dots, t_n\}$, where $t_i = i/n$. That is, g is the piecewise linear function such that $g(t_0) = 0$, $g(t_i) = -\sum_{k=1}^i \log \alpha_k$, $i = 1, \dots, n$. Then, define $\beta = g'$. This leads to

$$\alpha_i = e^{-\int_{t_{i-1}}^{t_i} \beta_r dr}, \quad i = 1, \dots, n,$$

and so

$$\bar{\alpha}_i = e^{-\int_0^{t_i} \beta_r dr}, \quad i = 1, \dots, n.$$

Note that since $-\log \alpha_i > 0$ the function β is nonnegative. Further, by (H2),

$$\lim_{n \rightarrow \infty} \int_0^1 \beta_t^{(n)} dt = \infty.$$

Let $\mathbb{F} = \{\mathcal{F}_t\}_{0 \leq t \leq 1}$ be a filtration with the usual conditions, i.e., $\mathcal{F}_t = \bigcap_{u > t} \mathcal{F}_u$ and $\mathcal{F}_0 \supset \mathcal{N}$, where \mathcal{N} denotes the collection of \mathbb{P} -null subsets from \mathcal{F} . Let $\{W_t\}_{t \geq 0}$ be a d -dimensional \mathbb{F} -Brownian motion. Then there exists a unique strong solution $X = \{X_t\}_{0 \leq t \leq 1}$ of the SDE

$$dX_t = -\frac{1}{2}\beta_t X_t dt + \sqrt{\beta_t} dW_t, \quad X_0 = \mathbf{x}_0.$$

Denote by $p(t, x, r, y)$ the transition density of $\{X_t\}$, i.e.,

$$(6) \quad p(t, x, r, y) = \frac{1}{(2\pi\sigma_{t,r}^2)^{d/2}} \exp\left(-\frac{|y - m_{t,r}x|^2}{2\sigma_{t,r}^2}\right), \quad 0 < t < r, \quad x, y \in \mathbb{R}^d,$$

where $m_{t,r} = e^{-\frac{1}{2} \int_t^r \beta_u du}$ and $\sigma_{t,r} = \sqrt{1 - m_{t,r}^2}$. The solution X_t is represented as

$$X_t = X_0 e^{-\frac{1}{2} \int_0^t \beta_r dr} + \int_0^t \sqrt{\beta_r} e^{-\frac{1}{2} \int_r^t \beta_u du} dW_r, \quad 0 \leq t \leq 1.$$

In particular, for any fixed i ,

$$(7) \quad X_{t_i} \sim \sqrt{\bar{\alpha}_i} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_i} Z_i.$$

Further, the density function $p_t(y) = p_t^{(n)}(y)$ of X_t is given by

$$p_t(y) := \int_{\mathbb{R}^d} p(0, x, t, y) \mu_{data}(dx), \quad t > 0, \quad y \in \mathbb{R}^d.$$

It is straightforward to check that the distribution of X_1 converges to the standard normal distribution. Precisely, we have

$$\lim_{n \rightarrow \infty} p_1^{(n)}(y) = \phi(y) := N(y; 0, I_d) = \frac{e^{-|y|^2/2}}{(2\pi)^{d/2}}, \quad y \in \mathbb{R}^d.$$

Further, p_t satisfies the forward Kolmogorov equation

$$(8) \quad \partial_t p_t(y) = \frac{1}{2} \beta_t \sum_{i=1}^d \partial_{y_i} (y_i p_t(y)) + \frac{\beta_t}{2} \Delta p_t(y), \quad t \in (t_i, t_{i+1}), \quad i = 0, \dots, n-1,$$

where Δ denotes the Laplacian with respect to the spatial variable.

As in (3), for $t \geq 0$ and $x \in \mathbb{R}^d$ we define $\nabla \log p_t(x)$ by

$$\nabla \log p_t(x) = \begin{cases} \nabla p_t(x)/p_t(x), & \text{if } p_t(x) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The condition (H1) means that the score function $\nabla \log p_t(x)$ has linear growth.

Lemma 2. *The function $\nabla \log p_t(x)$ satisfies*

$$|\nabla \log p_t(x)| \leq \frac{c_0}{m_{0,t}} + \frac{c_1}{m_{0,t}^2} |x|, \quad a.e. \ x \in \mathbb{R}^d, \quad 0 \leq t \leq 1.$$

Notice that by continuity, for $t > 0$ the inequality in Lemma 2 holds for any $x \in \mathbb{R}^d$. Thus Lemma 2 is a generalization of Lemma 1.

Let $\bar{X}_t = X_{1-t}$ for $t \in [0, 1]$. Then, by Lemma 2,

$$\mathbb{E} \int_0^1 \beta_t |\nabla \log p_t(X_t)| dt < \infty.$$

This together with Theorem 2.1 in [9] means that there exists a d -dimensional $\bar{\mathbb{F}}$ -Brownian motion $\{\bar{W}_t\}_{0 \leq t \leq 1}$ such that

$$(9) \quad d\bar{X}_t = \left[\frac{1}{2} \beta_{1-t} \bar{X}_t + \beta_{1-t} \nabla \log p_{1-t}(\bar{X}_t) \right] dt + \sqrt{\beta_{1-t}} d\bar{W}_t$$

where $\bar{\mathbb{F}} = \{\bar{\mathcal{F}}_t\}_{0 \leq t \leq 1}$ with $\bar{\mathcal{F}}_t = \sigma(\bar{X}_u : u \leq t) \vee \mathcal{N}$.

The following is a first key result, obtained by a generalized Girsanov–Maruyama theorem as stated in Liptser and Shiryaev [22, Chapter 6].

Lemma 3. *There exists a weak solution of the SDE*

$$(10) \quad dX_t^* = \left[\frac{1}{2} \beta_{1-t} X_t^* + \beta_{1-t} \nabla \log p_{1-t}(X_t^*) \right] dt + \sqrt{\beta_{1-t}} dW_t, \quad 0 \leq t \leq 1,$$

with initial condition $X_0^* \sim N(0, I_d)$. More precisely, there exist a filtration \mathbb{F}^* on (Ω, \mathcal{F}) , a probability measure \mathbb{P}^* on (Ω, \mathcal{F}) , an \mathbb{F}^* -Brownian motion $\{W_t^*\}_{0 \leq t \leq 1}$ under \mathbb{P}^* , and a continuous \mathbb{F}^* -adapted process $\{X_t^*\}_{0 \leq t \leq 1}$ such that $\{X_t^*\}$ satisfies the SDE (10) with $\{W_t\}$ replaced by $\{W_t^*\}$ such that $X_0^* \sim N(0, I_d)$ under \mathbb{P}^* . Further, the transition probability density $p^*(t, x, r, y)$ of X^* under \mathbb{P}^* is given by

$$p^*(t, x, r, y) = e^{\frac{d}{2} \int_t^r \beta_{1-u} du} \frac{p_{1-r}(y)}{p_{1-t}(x)} q(t, x, r, y), \quad 0 < t < r < 1, \quad x, y \in \mathbb{R}^d,$$

where

$$(11) \quad q(t, x, r, y) = \frac{m_{1-r, 1-t}^d}{(2\pi\sigma_{1-r, 1-t}^2)^{d/2}} \exp \left(-\frac{m_{1-r, 1-t}^2}{2\sigma_{1-r, 1-t}^2} \left| y - \frac{1}{m_{1-r, 1-t}} x \right|^2 \right).$$

The following lemma provides moment estimates that plays a key role in the subsequent argument.

Lemma 4. *We have*

$$\mathbb{E}^* |X_t^*|^2 \leq C d (\bar{\alpha}_n)^{-1/2} e^{2(c_0 + c_1)(\bar{\alpha}_n)^{-1}}$$

and

$$\mathbb{E}^* |X_t^*|^4 \leq C d^2 (\bar{\alpha}_n)^{-3} e^{2(4c_0 + 3c_1)(\bar{\alpha}_n)^{-1}}.$$

Next, we introduce the function s defined by for $t \in (t_{i-1}, t_i]$ with $i = 1, \dots, n$,

$$s(t, x) = -\frac{1 + \sqrt{\alpha_i}}{2\sqrt{1 - \alpha_i}} z_i(x), \quad x \in \mathbb{R}^d,$$

and $s(0, x) = 0$. Define $\hat{X}_0 = X_0^*$. For $i = 0, 1, \dots, n-1$, with given \hat{X}_{t_i} , there exists a unique strong solution $\{\hat{X}_t\}_{t_i \leq t \leq t_{i+1}}$ of the SDE

$$d\hat{X}_t = \left[\frac{1}{2} \beta_{1-t} \hat{X}_t + \beta_{1-t} s(1 - t_i, \hat{X}_{t_i}) \right] dt + \sqrt{\beta_{1-t}} dW_t^*$$

on $(\Omega, \mathcal{F}, \mathbb{F}^*, \mathbb{P}^*)$. Thus, \hat{X}_t satisfies

$$d\hat{X}_t = \left[\frac{1}{2} \beta_{1-t} \hat{X}_t + \beta_{1-t} s(1 - \tau_n(t), \hat{X}_{\tau_n(t)}) \right] dt + \sqrt{\beta_{1-t}} dW_t^*, \quad 0 \leq t \leq 1$$

with initial condition $\hat{X}_0 = X_0^*$, where $\tau_n(t)$ is such that $n\tau_n(t)$ is greatest integer not exceeding nt . Moreover, on $[t_j, t_{j+1}]$,

$$\hat{X}_t = e^{\frac{1}{2} \int_{t_j}^t \beta_{1-r} dr} \hat{X}_{t_j} + \int_{t_j}^t \beta_{1-r} e^{\frac{1}{2} \int_r^t \beta_{1-u} du} s(1 - t_j, \hat{X}_{t_j}) dr + \int_{t_j}^t \sqrt{\beta_{1-r}} e^{\frac{1}{2} \int_r^t \beta_{1-u} du} dW_r^*.$$

In particular,

$$(12) \quad \hat{X}_{t_{j+1}} = \frac{1}{\sqrt{\alpha_{n-j}}} \hat{X}_{t_j} + 2s(1 - t_j, \hat{X}_{t_j}) \frac{1 - \sqrt{\alpha_{n-j}}}{\sqrt{\alpha_{n-j}}} + \sqrt{\frac{1 - \alpha_{n-j}}{\alpha_{n-j}}} \hat{\xi}_{j+1},$$

where $\{\hat{\xi}_i\}_{i=1}^n$ is an IID sequence with common distribution $N(0, I_d)$ under \mathbb{P}^* . The process $\{\hat{X}_{t_i}\}_{i=0}^n$ can be seen as an exponential integrator type approximation of $\{X_t^*\}_{0 \leq t \leq 1}$ that appears in continuous time formulation (see [3], [7], [15], and [16]). Since

$$2s(t_i, x)(1 - \sqrt{\alpha_i}) = -\frac{1 - \alpha_i}{\sqrt{1 - \alpha_i}} z_i(x),$$

setting $j = n - i$ in (12), we have

$$(13) \quad \mathbb{P}^*(\hat{X}_{t_{n-i}})^{-1} = \mathbb{P}(\mathbf{x}_i^*)^{-1}, \quad i = 1, \dots, n.$$

Denote by $D_{KL}(\mu \| \nu)$ the Kullback-Leibler divergence or the relative entropy of $\mu \in \mathcal{P}(\mathbb{R}^d)$ with respect to $\nu \in \mathcal{P}(\mathbb{R}^d)$. Let $p_t^*(x)$ be the density of X_t^* under \mathbb{P}^* . The theory of Schrödinger bridges provides a way to estimate the reverse-time distributional error $D_{TV}(\mu_{data}, p_1^*(x)dx)$ in terms of the forward-time one $D_{KL}(\phi(x)dx \| p_1(x)dx)$.

Lemma 5. *We have*

$$D_{TV}(\mu_{data}, p_1^*(x)dx) \leq \sqrt{-\frac{1}{2} D_{KL}(\phi(x)dx \| p_1(x)dx) + \frac{m_{0,1}^2 + m_{0,1}}{4(1 - m_{0,1}^2)} (d + \mathbb{E}|\hat{\mathbf{x}}_0|^2)}.$$

Denote by \mathbb{E}^* the expectation under \mathbb{P}^* . By combining Pinsker's inequality and Girsanov-Maruyama theorem, we see the following:

Lemma 6. *We have*

$$(14) \quad D_{TV}(\mathbb{P}^*(\hat{X}_1)^{-1}, \mathbb{P}^*(X_1^*)^{-1}) \leq \frac{1}{2} \sqrt{\mathbb{E}^* \int_0^1 \beta_{1-t} |s(1-t, X_t^*) - \nabla \log p_{1-t}(X_t^*)|^2 dt}.$$

As for the right-hand side in (14), we have

$$(15) \quad \begin{aligned} & \mathbb{E}^* \int_0^1 \beta_{1-t} |s(1-t, X_t^*) - \nabla \log p_{1-t}(X_t^*)|^2 dt \\ & \leq 2 \sum_{i=0}^{n-1} \mathbb{E}^* |s(1-t_i, X_{t_i}^*) - \nabla \log p_{1-t_i}(X_{t_i}^*)|^2 \int_{t_i}^{t_{i+1}} \beta_{1-t} dt \\ & \quad + 2 \mathbb{E}^* \int_0^1 \beta_{1-t} \left| \nabla \log p_{1-\tau_n(t)}(X_{\tau_n(t)}^*) - \nabla \log p_{1-t}(X_t^*) \right|^2 dt. \end{aligned}$$

To estimate the first term of the right-hand side in (15), we start with observing a relation of $\mathbb{E}|s(t, X_t) - \nabla \log p_t(X_t)|^2$ and the noise estimating objective. For a fixed i we have

$$\begin{aligned} \mathbb{E} \mathbf{s}_i(X_{t_i})^\top \nabla \log p_{t_i}(X_{t_i}) &= \int_{\mathbb{R}^d} \mathbf{s}_i(y)^\top (\nabla \log p_{t_i}(y)) p_{t_i}(y) dy \\ &= \int_{\mathbb{R}^d} \mathbf{s}_i(y)^\top \nabla \left(\int_{\mathbb{R}^d} p(0, x, t_i, y) \mu_{data}(dx) \right) dy \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \mathbf{s}_i(y)^\top \frac{\nabla_y p(0, x, t_i, y)}{p(0, x, t_i, y)} p(0, x, t_i, y) dy \mu_{data}(dx) \\ &= \int_{\mathbb{R}^d} \mathbb{E} \left[\mathbf{s}_i(X_{t_i})^\top \nabla_y \log p(0, x, t_i, X_{t_i}) \middle| X_0 = x \right] \mu_{data}(dx) \\ &= \mathbb{E} \left[\mathbf{s}_i(X_{t_i})^\top \nabla_y \log p(0, X_0, t_i, X_{t_i}) \right], \end{aligned}$$

where for simplicity we have denoted $\nabla_y \log p(0, X_0, t_i, X_{t_i}) = \nabla_y \log p(0, X_0, t_i, y)|_{y=X_{t_i}}$. Thus

$$\begin{aligned} & \mathbb{E} |\mathbf{s}_i(X_{t_i}) - \nabla \log p_{t_i}(X_{t_i})|^2 \\ &= \mathbb{E} |\mathbf{s}_i(X_{t_i}) - \nabla_y \log p(0, X_0, t_i, X_{t_i})|^2 + \mathbb{E} [|\nabla \log p_{t_i}(X_{t_i})|^2 - |\nabla_y \log p(0, X_0, t_i, X_{t_i})|^2]. \end{aligned}$$

Using (6), we get

$$\nabla_y \log p(0, X_0, t_i, X_{t_i}) = -\frac{1}{\sigma_{0,t_i}^2} (X_{t_i} - m_{0,t_i} X_0) \sim -\frac{1}{\sqrt{1-\bar{\alpha}_i}} Z_i.$$

This together with definition of \mathbf{s}_i and (7) leads to

$$\mathbb{E} |\mathbf{s}_i(X_{t_i}) - \nabla_y \log p(0, X_0, t_i, X_{t_i})|^2 = \mathbb{E} |\mathbf{s}_i(\mathbf{x}_i) - \nabla \log \mathbf{p}_i(\mathbf{x}_i | \mathbf{x}_0)|^2 = \frac{1}{1-\bar{\alpha}_i} \mathbb{E} |z_i(\mathbf{x}_i) - Z_i|^2,$$

whence (2) follows.

The following lemma is obtained by estimating $\mathbb{E}^* |s(1-t, X_t^*) - \nabla \log p_{1-t}(X_t^*)|^2$ in terms of $\mathbb{E} |s(1-t, X_{1-t}) - \nabla \log p_{1-t}(X_{1-t})|^2$:

Lemma 7. *There exists $\delta_1 \in (0, 1/4)$ such that If $\bar{\alpha}_n < \delta_1$ then*

$$\sum_{i=0}^{n-1} \mathbb{E}^* |s(1 - t_i, X_{t_i}^*) - \nabla \log p_{1-t_i}(X_{t_i}^*)|^2 \int_{t_i}^{t_{i+1}} \beta_{1-t} dt \leq C(-n \log \alpha_{\min})(\bar{\alpha}_n)^{-2} \sqrt{dL}.$$

To estimate the second term of the right-hand side in (15) we characterize the process

$$Y_t^* := \nabla \log p_{1-t}(X_t^*), \quad 0 \leq t \leq 1,$$

as one of components of a solution of a forward-backward SDE.

Lemma 8. *There exist $\delta_2 \in (0, \delta_1]$ and an $\mathbb{R}^{d \times d}$ -valued, continuous, and adapted process $\{Z_t^*\}_{0 \leq t \leq 1}$ such that if $\bar{\alpha}_n < \delta_2$ then*

$$(16) \quad \mathbb{E}^* \int_{t_1}^{t_2} |Z_t^*|^2 dt \leq C d^2 (\bar{\alpha}_n)^{-10} e^{(10c_0 + 7c_1)(\bar{\alpha}_n)^{-1}} \left(e^{\int_0^{1-t_1} \beta_r dr} - e^{\int_0^{1-t_2} \beta_r dr} \right), \quad 0 \leq t_1 < t_2 \leq 1,$$

where $|A|$ stands for the Frobenius norm of $A \in \mathbb{R}^{d \times d}$, and the triple (X^*, Y^*, Z^*) solves the following forward-backward SDE: for $0 \leq t \leq 1$,

$$(17) \quad \begin{aligned} X_t^* &= X_0 + \int_0^t \beta_{1-r}(X_r^* + Y_r^*) ds + \int_0^t \sqrt{\beta_{1-r}} dW_r^*, \\ \nabla \log p_{\text{data}}(X_1^*) &= Y_t^* - \frac{1}{2} \int_t^1 \beta_{1-r} Y_r^* dr + \int_t^1 Z_r^* dW_r^*. \end{aligned}$$

We are now ready to prove our main theorem.

Proof of Theorem 1. Step (i). Put $\sigma = \sigma_{0,1}$ and $m = m_{0,1}$ for notational simplicity. Observe

$$\frac{m^2 + m}{2(1 - m^2)}(d + |y|^2) \leq 2m(d + |y|^2)$$

if $m^2 = \bar{\alpha}_n \leq 1/2$. This together with Lemma 5 yields

$$(18) \quad D_{TV}(\mu_{\text{data}}, p_1^*(x) dx) \leq C d^{1/2} \bar{\alpha}_n^{1/4}.$$

Step (ii). Lemma 7 tells us that if $\bar{\alpha}_n < \delta_1$ then

$$\sum_{i=0}^{n-1} \mathbb{E}^* |s(1 - t_i, X_{t_i}^*) - \nabla \log p_{1-t_i}(X_{t_i}^*)|^2 \int_{t_i}^{t_{i+1}} \beta_{1-t} dt \leq C \sqrt{d} (-n \log \alpha_{\min})(\bar{\alpha}_n)^{-2} \sqrt{L}.$$

To estimate the second term of the right-hand side in (15), observe

$$(19) \quad \mathbb{E}^* \int_0^1 \beta_{1-t} \left| \nabla \log p_{1-\tau_n(t)}(X_{\tau_n(t)}^*) - \nabla \log p_{1-t}(X_t^*) \right|^2 dt = \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} \beta_{1-t} \mathbb{E}^* |Y_{t_i}^* - Y_t^*|^2 dt.$$

For $t \in [t_i, t_{i+1})$, $i = 0, 1, \dots, n-1$, by Lemma 8,

$$\mathbb{E}^* |Y_{t_i}^* - Y_t^*|^2 \leq \frac{1}{2} \mathbb{E}^* \left| \int_{t_i}^t \beta_{1-r} Y_r^* dt \right|^2 + 2 \int_{t_i}^t \mathbb{E}^* |Z_r^*|^2 dr \leq \frac{1}{2n} \int_{t_i}^t \beta_{1-r}^2 \mathbb{E}^* |Y_r^*|^2 dr + 2 \int_{t_i}^t \mathbb{E}^* |Z_r^*|^2 dr.$$

By Lemmas 2 and 4,

$$\mathbb{E}^* |Y_t^*|^2 \leq \frac{2c_0^2}{m_{0,1-r}^2} + \frac{2c_1^2}{m_{0,1-r}^4} \mathbb{E}^* |X_r^*|^2 \leq \frac{2c_0^2}{m_{0,1-r}^2} + \frac{Cd}{m_{0,1-r}^4} (\bar{\alpha}_n)^{-3/2} e^{2(c_0+c_1)(\bar{\alpha}_n)^{-1}}$$

and so

$$\begin{aligned} \int_{t_i}^{t_{i+1}} \beta_{1-r}^2 \mathbb{E}^* |Y_r^*|^2 dr &\leq C(-n \log \alpha_{n-i}) \int_{t_i}^{t_{i+1}} \beta_{1-r} e^{\int_0^{1-r} \beta_u du} dr \\ &\quad + Cd(-n \log \alpha_{n-i}) (\bar{\alpha}_n)^{-3/2} e^{2(c_0+c_1)(\bar{\alpha}_n)^{-1}} \int_{t_i}^{t_{i+1}} \beta_{1-r} e^{2 \int_0^{1-r} \beta_u du} dr \\ &\leq Cd(-n \log \alpha_{min}) (\bar{\alpha}_n)^{-3/2} e^{2(c_0+c_1)(\bar{\alpha}_n)^{-1}} \left(e^{\int_0^{1-t_i} \beta_r dr} - e^{\int_0^{1-t_{i+1}} \beta_r dr} \right). \end{aligned}$$

Further, by Lemma 8, if $\bar{\alpha}_n < \delta_2 \leq \delta_1$ then

$$\int_{t_i}^{t_{i+1}} \mathbb{E}^* |Z_r^*|^2 dr \leq Cd^2 (\bar{\alpha}_n)^{-10} e^{(10c_0+7c_1)(\bar{\alpha}_n)^{-1}} \left(e^{\int_0^{1-t_i} \beta_r dr} - e^{\int_0^{1-t_{i+1}} \beta_r dr} \right).$$

Therefore, there exists $\delta \in (0, \delta_2)$ such that if $\bar{\alpha}_n < \delta_3$ then the right-hand side in (19) is at most

$$Cd^2 n (\log \alpha_{min})^2 e^{(10c_0+7c_1+1)(\bar{\alpha}_n)^{-1}}.$$

Step (iii). We have

$$D_{TV}(\mathbb{P}(\mathbf{x}_0^*)^{-1}, \mu_{data}) \leq A_1 + A_2,$$

where $A_1 = D_{TV}(\mathbb{P}^*(X_1^*)^{-1}, \mu_{data})$ and $A_2 = D_{TV}(\mathbb{P}^*(\hat{X}_1)^{-1}, \mathbb{P}^*(X_1^*)^{-1})$. It follows from Steps (i) and (ii) that if $\bar{\alpha}_n < \delta$ then $A_1^2 \leq Cd\sqrt{\bar{\alpha}_n}$ and

$$A_2^2 \leq C\sqrt{d}(-n \log \alpha_{min}) (\bar{\alpha}_n)^{-2} \sqrt{L} + Cd^2 e^{(10c_0+7c_1+1)(\bar{\alpha}_n)^{-1}} n (\log \alpha_{min})^2.$$

Thus Theorem 1 follows. □

A proof of Corollary 1 is elementary, so omitted.

3 Proofs of the lemmas

This section is devoted to proofs of Lemmas 1–8.

Proof of Lemma 2. Fix $t \in (0, 1]$ and put $\sigma = \sigma_{0,t}$ and $m = m_{0,t}$ for notational simplicity. Using

$$(20) \quad \partial_{y_k} e^{-\frac{|y-mx|^2}{2\sigma^2}} = -\frac{y_k - mx_k}{\sigma^2} e^{-\frac{|y-mx|^2}{2\sigma^2}} = -\frac{1}{m} \partial_{x_k} e^{-\frac{|y-mx|^2}{2\sigma^2}}$$

and the integration-by-parts formula, we find

$$\nabla p_t(y) = -\frac{1}{m} \int_{\mathbb{R}^d} \nabla_x p(0, x, t, y) p_{data}(x) dx = \frac{1}{m} \int_{\mathbb{R}^d} p(0, x, t, y) \nabla p_{data}(x) dx.$$

So, again by (20)

$$\begin{aligned} \nabla p_t(y) &= \frac{1}{m} \int_{\mathbb{R}^d} p(0, x, t, y) (\nabla p_{data}(x) + p_{data}(x) Q x) dx \\ &\quad - \frac{1}{m} Q \int_{\mathbb{R}^d} \left(\frac{\sigma^2}{m} \nabla_y p(0, x, t, y) + \frac{1}{m} p(0, x, t, y) y \right) p_{data}(x) dx, \end{aligned}$$

whence by (H1),

$$\left| \left(I_d + \frac{\sigma^2}{m^2} Q \right) \nabla p_t(y) \right| \leq \frac{c_0}{m} p_t(y) + \frac{c_1}{m^2} |y| p_t(y).$$

Hence, for any $t > 0$ and $y \in \mathbb{R}^d$,

$$|\nabla p_t(y)| = \left| \left(I_d + \frac{\sigma^2}{m^2} Q \right)^{-1} \left(I_d + \frac{\sigma^2}{m^2} Q \right) \nabla p_t(y) \right| \leq \frac{c_0}{m} p_t(y) + \frac{c_1}{m^2} |y| p_t(y).$$

For $t = 0$ the condition (H1) directly leads to

$$|\nabla p_0(y)| \leq |\nabla p_0(y) + Q y p_0(y)| + |Q y| p_0(y) \leq c_0 p_0(y) + c_1 |y| p_0(y), \quad \text{a.e. } y \in \mathbb{R}^d.$$

Thus the lemma follows. \square

Proof of Lemma 3. Let $\eta \sim N(0, I_d)$ under \mathbb{P} and be independent of X . Define the filtration $\mathbb{F}^* = \{\mathcal{F}_t^*\}_{0 \leq t \leq 1}$ by $\mathcal{F}_t^* = \sigma(\bar{\mathcal{F}}_t \cup \sigma(\eta))$, $0 \leq t \leq 1$. Note that \bar{W} is an $(\mathbb{F}^*, \mathbb{P})$ -Brownian motion. Let $\{Y_t\}_{0 \leq t \leq 1}$ be a unique strong solution of

$$dY_t = \frac{1}{2} \beta_{1-t} Y_t dt + \sqrt{\beta_{1-t}} d\bar{W}_t, \quad Y_0 = \eta$$

on $(\Omega, \mathcal{F}, \mathbb{F}^*, \mathbb{P})$. Let

$$Y_r^{t,x} = e^{\frac{1}{2} \int_t^r \beta_{1-u} du} x + \int_t^r \sqrt{\beta_{1-u}} e^{\frac{1}{2} \int_t^u \beta_{1-\tau} d\tau} d\bar{W}_u.$$

Then the mean vector and the covariance matrix of $Y_r^{t,x}$ is given respectively by

$$e^{\frac{1}{2} \int_{1-r}^{1-t} \beta_u du} x = \frac{1}{m_{1-r,1-t}} x,$$

and

$$\left(e^{\int_{1-r}^{1-t} \beta_u du} - 1 \right) I_d = \frac{\sigma_{1-r,1-t}^2}{m_{1-r,1-t}^2} I_d.$$

Thus the transition density of $\{Y_t\}$ is given by $q(t, x, r, y)$ as in (11).

Now, put

$$h(t, y) = e^{\frac{d}{2} \int_0^t \beta_{1-u} du} p_{1-t}(y), \quad 0 \leq t \leq 1, \quad y \in \mathbb{R}^d.$$

Then a simple application of Itô formula yields

$$dh(t, Y_t) = \sqrt{\beta_{1-t}} e^{\frac{d}{2} \int_0^t \beta_{1-u} du} \nabla p_{1-t}(Y_t) d\bar{W}_t.$$

The condition (H1) means that $\{h(t, Y_t)\}_{0 \leq t \leq 1}$ is an $(\mathbb{F}^*, \mathbb{P})$ -martingale. Since $h \geq 0$, the conditional expectation $\mathbb{E}[h(1, Y_1)/h(0, Y_0) | \mathcal{F}_0]$ exists and equal to $\mathbb{E}[h(1, Y_1) | \mathcal{F}_0]/h(0, Y_0) = 1$, whence $\mathbb{E}[h(1, Y_1)/h(0, Y_0)] = 1$. Thus, by a generalized Girsanov–Maruyama theorem (see [22, Theorem 6.2]), the process

$$W_t^* := \bar{W}_t - \int_0^t \sqrt{\beta_{1-r}} \nabla \log p_{1-r}(Y_r) dr, \quad 0 \leq t \leq 1,$$

is an \mathbb{F}^* -Brownian motion under the probability measure \mathbb{P}^* defined by $d\mathbb{P}^*/d\mathbb{P} = h(1, Y_1)/h(0, Y_0)$. Furthermore, $\{Y_t\}$ satisfies (10) with W replaced by W^* . Hence $(\Omega, \mathcal{F}, \mathbb{F}^*, \mathbb{P}^*, W^*, Y)$ is a weak solution of (10).

To derive the representation of the transition density, take arbitrary $A \in \mathcal{B}(\mathbb{R}^d)$, $t < r$ and observe

$$\mathbb{P}^*(Y_r \in A | \mathcal{F}_t^*) = \frac{1}{h(t, Y_t)} \mathbb{E}[1_{\{Y_r \in A\}} h(r, Y_r) | \mathcal{F}_t^*] = \int_A \frac{h(r, y)}{h(t, Y_t)} q(t, Y_t, r, y) dy.$$

Thus the lemma follows. □

Proof of Lemma 4. Applying the Itô formula for $|X_t^*|^2$ and then using Lemma 2, we get

$$\begin{aligned} \mathbb{E}^* |X_t^*|^2 &= \mathbb{E}^* |X_0^*|^2 + \int_0^t \beta_{1-u} \mathbb{E}^* \left[|X_u^*|^2 + 2(X_u^*)^\top \nabla \log p_{1-u}(X_u^*) + d \right] du \\ &\leq d + \int_0^t \beta_{1-u} \mathbb{E}^* \left[|X_u^*|^2 + \frac{2c_0 |X_u^*|}{m_{0,1-u}} + \frac{2c_1 |X_u^*|^2}{m_{0,1-u}^2} + d \right] du \\ &\leq d + \int_0^t \beta_{1-u} \left\{ \left(1 + \frac{c_0}{m_{0,1-u}} + \frac{2c_1}{m_{0,1-u}^2} \right) \mathbb{E}^* |X_u^*|^2 + \frac{c_0}{m_{0,1-u}} + d \right\} du. \end{aligned}$$

Thus Gronwall's inequality yields

$$\mathbb{E}^* |X_t^*|^2 \leq \left(d + \int_0^1 \beta_{1-u} (c_0 m_{0,1-u}^{-1} + d) du \right) \exp \left(\int_0^t \beta_{1-u} (1 + c_0 m_{0,1-u}^{-1} + 2c_1 m_{0,1-u}^{-2}) du \right).$$

Observe

$$\int_0^t \beta_{1-u} m_{0,1-u}^{-1} du \leq 2(\bar{\alpha}_n)^{-1/2}, \quad \int_0^t \beta_{1-u} m_{0,1-u}^{-2} du \leq (\bar{\alpha}_n)^{-1}$$

and

$$\int_0^t \beta_{1-u} du \leq -\log \bar{\alpha}_n = -2 \log \sqrt{\bar{\alpha}_n} \leq 2(\bar{\alpha}_n)^{-1/2} - 2.$$

So we get

$$\mathbb{E}^*|X_t^*|^2 \leq Cd(\bar{\alpha}_n)^{-1/2}e^{2(c_0+c_1)(\bar{\alpha}_n)^{-1}}.$$

Next, applying the Itô formula for $|X_t^*|^4$, we get

$$\begin{aligned} \mathbb{E}^*|X_t^*|^4 &= \mathbb{E}^*|X_0^*|^4 + 4\mathbb{E}^* \int_0^t |X_u^*|^2 \left\{ \frac{1}{2}\beta_{1-u}|X_u^*|^2 + \beta_{1-u}(X_u^*)^\top \nabla \log p_{1-u}(X_u^*) + \frac{d+2}{2}\beta_{1-u} \right\} du \\ &\leq d(d+2) + \int_0^t \beta_{1-u} \mathbb{E}^* \left[2|X_u^*|^4 + 4|X_u^*|^3 \left(c_0 m_{0,1-u}^{-1} + c_1 m_{0,1-u}^{-2} |X_u^*| \right) + (2d+4)|X_u^*|^2 \right] du \\ &\leq d(d+2) + \int_0^t \beta_{1-u} \left(2 + 3c_0 m_{0,1-u}^{-1} + 4c_1 m_{0,1-u}^{-2} \right) \mathbb{E}^*|X_u^*|^4 du \\ &\quad + Cd^2(\bar{\alpha}_n)^{-1/2}e^{2(c_0+c_1)(\bar{\alpha}_n)^{-1}} \int_0^t \beta_{1-u} du, \end{aligned}$$

where we have used Young's inequality $4a^3b \leq 3a^4 + b^4$ for $a, b \in \mathbb{R}$ and the estimate of $\mathbb{E}^*|X_t^*|^2$ obtained just above. Then by Gronwall's inequality and $-\log \eta \leq 2\eta^{-1/2}$ for $\eta \in (0, 1]$,

$$\begin{aligned} \mathbb{E}^*|X_t^*|^4 &\leq Cd^2 \left\{ 1 + (\bar{\alpha}_n)^{-1}e^{2(c_0+c_1)(\bar{\alpha}_n)^{-1}}(-\log \bar{\alpha}_n) \right\} \exp \left(\int_0^t \beta_{1-u} (2 + 3c_0 m_{0,1-u}^{-1} + 4c_1 m_{0,1-u}^{-2}) du \right) \\ &\leq Cd^2(\bar{\alpha}_n)^{-3}e^{(8c_0+6c_1)(\bar{\alpha}_n)^{-1}}. \end{aligned}$$

Thus the lemma follows. \square

Proof of Lemma 5. Let $P_1 = \mathbb{P}^*(X_1^*)^{-1} = p_1^*(x)dx$ and $P_{01} = \mathbb{P}^*(X_0^*, X_1^*)^{-1}$. Then consider the Schrödinger's bridge problem

$$(21) \quad \inf D_{KL}(R \| P_{01}),$$

where the infimum is taken over all $R \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ such that $R(dx \times \mathbb{R}^d) = \phi(x)dx$ and $R(\mathbb{R}^d \times dx) = \mu_{data}(dx)$. It is known that this problem has a unique minimizer R^* , provided that (21) is finite (see, e.g., Rüschendorf and Thomsen [32]). To confirm this, we shall pick up the product measure $\phi(x)dx \otimes \mu_{data}$ to see

$$\begin{aligned} &\log p_{data}(x) - \log p^*(0, y, 1, x) \\ &= -\log \frac{\phi(x)}{p_1(x)} + \log \sigma_{0,1}^d + \frac{1}{2(1-m_{0,1}^2)}(m_{0,1}^2|x|^2 - 2m_{0,1}x^\top y + m_{0,1}^2|y|^2) \\ &\leq -\log \frac{\phi(x)}{p_1(x)} + \frac{m_{0,1}^2 + m_{0,1}}{2(1-m_{0,1}^2)}(|x|^2 + |y|^2). \end{aligned}$$

Thus

$$\begin{aligned} \int_{\mathbb{R}^d} D_{KL}(\phi(x)dx \| p^*(0, y, 1, x)dx) p_{data}(y)dy &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} (\log \phi(x) - \log p^*(0, y, 1, x)) \phi(x)dx p_{data}(y)dy \\ &\leq -D_{KL}(\phi(x)dx \| p_1(x)dx) + \frac{m_{0,1}^2 + m_{0,1}}{2(1-m_{0,1}^2)}(d + \mathbb{E}|\hat{\mathbf{x}}_0|^2) \end{aligned}$$

whence

$$\begin{aligned} D_{KL}(\phi(x)dx \otimes \mu_{data} \parallel P_{01}) &= \int_{\mathbb{R}^d \times \mathbb{R}^d} \log \frac{p_{data}(y)}{p^*(0, x, 1, y)} p_{data}(y) \phi(x) dx dy \\ &= \int_{\mathbb{R}^d} D_{KL}(\phi(x)dx \parallel p^*(0, x, 1, y)dx) p_{data}(y) dy < \infty, \end{aligned}$$

Therefore, since the optimal P^* of course satisfies $R^*(\mathbb{R}^d \times dx) = \mu_{data}(dx)$, we obtain

$$\begin{aligned} D_{TV}(\mu_{data}, P_1) &\leq D_{TV}(R^*, P_{01}) \leq \sqrt{\frac{1}{2} D_{KL}(R^* \parallel P_{01})} \\ &\leq \sqrt{\frac{1}{2} \int_{\mathbb{R}^d} D_{KL}(\phi(x)dx \otimes \mu_{data} \parallel P_{01}) p_{data}(y) dy} \\ &\leq \sqrt{-\frac{1}{2} D_{KL}(\phi(x)dx \parallel p_1(x)dx) + \frac{m_{0,1}^2 + m_{0,1}}{4(1 - m_{0,1}^2)} (d + \mathbb{E}|\hat{\mathbf{x}}_0|^2)} \end{aligned}$$

where the second inequality follows from Pinsker's inequality (see, e.g., Tsybakov [38]). \square

Proof of Lemma 6. We shall borrow the arguments in the proof of Theorem 7.18 in [22]. Denote by \mathbb{W}^d the space of \mathbb{R}^d -valued continuous functions on $[0, 1]$. Put $\hat{P} = \mathbb{P}^*(\hat{X})^{-1} \in \mathcal{P}(\mathbb{W}^d)$ and $P^* = \mathbb{P}^*(X^*)^{-1} \in \mathcal{P}(\mathbb{W}^d)$. Define the function κ by $\kappa(t, w) = \nabla \log p_{1-t}(w_t) - s(1 - \tau_n(t), w_{\tau_n(t)})$ for $w = (w_t) \in \mathbb{W}^d$. For any $N \geq 1$ consider the stopping time

$$T_N(w) := \inf \left\{ t \geq 0 : \int_0^t \beta_{1-u} |\kappa(u, w)|^2 du \geq N \right\} \wedge 1.$$

Since the function s is piecewise constant, the equation

$$\begin{aligned} X_t^{*,N} &= X_{t \wedge T_N(X^*)}^* + \int_0^t 1_{\{T_N(X^*) < u\}} \left(\frac{1}{2} \beta_{1-u} X_u^{*,N} + \beta_{1-u} g_u(X^{*,N}) \right) du \\ &\quad + \int_0^t 1_{\{T_N(X^*) < u\}} \sqrt{\beta_{1-u}} dW_u^* \end{aligned}$$

has a unique strong solution $\{X_t^{*,N}\}_{0 \leq t \leq 1}$, where $g_t(w) = s(1 - \tau_n(t), w) + 1_{\{T_N(w) \geq t\}} \kappa(t, w)$. Note that $X^{*,N}$ satisfies $X_t^{*,N} = X_t^*$ on $\{t \leq T_N(X^*)\}$. It is straightforward to see

$$dX_t^{*,N} = \left(\frac{1}{2} \beta_{1-t} X_t^{*,N} + \beta_{1-t} g_t(X^{*,N}) \right) dt + \sqrt{\beta_{1-t}} dW_t^*.$$

Then consider the process

$$\widetilde{W}_t^N := W_t^* + \int_0^t \sqrt{\beta_{1-r}} (g_r(X^{*,N}) - s(1 - \tau_n(r), X_{\tau_n(r)}^{*,N})) dr.$$

By the definition of T_N ,

$$\int_0^1 \beta_{1-r} |g_r(X^{*,N}) - s(1 - \tau_n(r), X_r^{*,N})|^2 dr = \int_0^{T_N(X^*)} \beta_{1-r} |\kappa(r, X^*)|^2 dr \leq N,$$

whence Novikov's condition is satisfied. So we can apply Girsanov-Maruyama theorem to see that \widetilde{W}^N is a Brownian motion under $\widetilde{\mathbb{P}}^N$ defined by

$$\begin{aligned} \frac{d\widetilde{\mathbb{P}}^N}{d\mathbb{P}^*} &= \exp \left[- \int_0^1 \sqrt{\beta_{1-t}} (g_t(X^{*,N}) - s(1 - \tau_n(t), X_{\tau_n(t)}^{*,N}))^\top dW_t^* \right. \\ &\quad \left. - \frac{1}{2} \int_0^1 \beta_{1-t} |g_t(X^{*,N}) - s(1 - \tau_n(t), X_{\tau_n(t)}^{*,N})|^2 dt \right]. \end{aligned}$$

Then $X^{*,N}$ satisfies

$$dX_t^{*,N} = \left[\frac{1}{2} \beta_{1-t} X_t^{*,N} + \beta_{1-t} s(1 - \tau_n(t), X_{\tau_n(t)}^{*,N}) \right] dt + \sqrt{\beta_{1-t}} d\widetilde{W}_t^N.$$

By the strong uniqueness, we have $\hat{P} = \widetilde{\mathbb{P}}^N(X^{*,N})^{-1}$ for any N . Hence, for $\Gamma \in \mathcal{B}(\mathbb{W}^d)$,

$$\begin{aligned} \hat{P}(\Gamma) &= \lim_{N \rightarrow \infty} \hat{P}(\Gamma \cap \{T_N = 1\}) = \lim_{N \rightarrow \infty} \widetilde{\mathbb{P}}^N(X^{*,N} \in \Gamma \cap \{T_N = 1\}) \\ &= \lim_{N \rightarrow \infty} \mathbb{E}^* \left[1_{\{X^{*,N} \in \Gamma \cap \{T_N = 1\}\}} \frac{d\widetilde{\mathbb{P}}^N}{d\mathbb{P}^*} \right] \\ &= \lim_{N \rightarrow \infty} \mathbb{E}^* \left[1_{\{X^{*,N} \in \Gamma \cap \{T_N = 1\}\}} \right. \\ &\quad \times \exp \left(- \int_0^{T_N(X^{*,N})} \sqrt{\beta_{1-t}} \kappa(t, X^*)^\top dW_t^* - \frac{1}{2} \int_0^{T_N(X^{*,N})} \beta_{1-t} |\kappa(t, X^*)|^2 dt \right) \Big] \\ &= \lim_{N \rightarrow \infty} \mathbb{E}^* \left[1_{\{X^* \in \Gamma \cap \{T_N = 1\}\}} \exp \left(- \int_0^1 \sqrt{\beta_{1-t}} \kappa(t, X^*)^\top dW_t^* - \frac{1}{2} \int_0^1 \beta_{1-t} |\kappa(t, X^*)|^2 dt \right) \right] \\ &= \mathbb{E}^* \left[1_{\{X^* \in \Gamma\}} \exp \left(- \int_0^1 \sqrt{\beta_{1-t}} \kappa(t, X^*)^\top dW_t^* - \frac{1}{2} \int_0^1 \beta_{1-t} |\kappa(t, X^*)|^2 dt \right) \right]. \end{aligned}$$

Since $\{W_t^*\}$ is adapted to the augmented natural filtration \mathbb{G} generated by $\{X_t^*\}$ and $\{\kappa(t, X^*)\}_{0 \leq t \leq 1}$ is \mathbb{G} -adapted, as in the proof of Lemma 2.4 in [13], we have

$$\int_0^1 \sqrt{\beta_{1-t}} \kappa(t, X^*)^\top dW_t^* = \lim_{k \rightarrow \infty} \int_0^1 (\kappa_t^{(k)})^\top dW_t^*$$

holds almost surely along subsequence for some \mathbb{G} -adapted simple processes $\{\kappa_t^{(k)}\}_{0 \leq t \leq 1}$, $k \in \mathbb{N}$. Thus, there exists a $\mathcal{B}(\mathbb{W}^d)$ -measurable map Φ such that

$$\Phi(X^*) = \exp \left(\int_0^1 \sqrt{\beta_{1-t}} \kappa(t, X^*)^\top dW_t^* - \frac{1}{2} \int_0^1 \beta_{1-s} |\kappa(t, X^*)|^2 dt \right), \quad \mathbb{P}^*\text{-a.s.}$$

This means

$$(22) \quad \hat{P}(\Gamma) = \mathbb{E}^* [1_{\{X^* \in \Gamma\}} \Phi(X^*)], \quad \Gamma \in \mathcal{B}(\mathbb{W}^d).$$

Now, again by Pinsker's inequality,

$$D_{TV}(P^*, \hat{P})^2 \leq \frac{1}{2} D_{KL}(P^* \parallel \hat{P}),$$

where by abuse of notation we have denoted the total variation distance and the KL-divergence on $\mathcal{P}(\mathbb{W}^d)$ by D_{TV} and D_{KL} , respectively.

Lemma 4 means $\sup_{0 \leq t \leq 1} \mathbb{E}^* |X_t^*|^2 < \infty$. So again by the linear growth of κ the Itô integral $\int_0^t \sqrt{\beta_{1-u}} \kappa(u, X^*) dW_u^*$ is a \mathbb{P}^* -martingale. Hence using (22) we find

$$\begin{aligned} D_{KL}(P^* \parallel \hat{P}) &= \int_{\mathbb{W}^d} \log \frac{dP^*}{d\hat{P}} dP^* = \int_{\mathbb{W}^d} (-\log \Phi(w)) \mathbb{P}^*(X^*)^{-1}(dw) \\ &= \frac{1}{2} \mathbb{E}^* \int_0^1 \beta_{1-t} |\kappa(t, X^*)|^2 dt. \end{aligned}$$

Thus the lemma follows. \square

Proof of Lemma 7. Hereafter, we shall often write $\sigma = \sigma_{0,1}$ and $m = m_{0,1}$ for notational simplicity. Step (i). We shall confirm that under (H1)

$$(23) \quad \mathbb{E}[e^{c|\mathbf{x}_0|^2}] < \infty$$

for some $c > 0$.

Let $x_0 \in \mathbb{R}^d$ be fixed such that $\log p_{data}(x_0) > 0$. By Taylor's theorem and (H1), for almost every $x \in \mathbb{R}^d$ with $p_{data}(x) > 0$,

$$\begin{aligned} &\log p_{data}(x) - \log p_{data}(x_0) + \frac{1}{2}(x - x_0)^\top Q(x - x_0) \\ &= \int_0^1 (\nabla \log p_{data}(x_0 + t(x - x_0)) + Q(x_0 + t(x - x_0)))^\top (x - x_0) dt \\ &\leq c_0 |x - x_0|, \end{aligned}$$

whence

$$\log p_{data}(x) \leq \log p_{data}(x_0) - \frac{1}{2} \lambda_{min} |x - x_0|^2 + c_0 |x - x_0| \leq C_0 - \frac{1}{4} \lambda_{min} |x|^2,$$

where $\lambda_{min} > 0$ is the minimum eigenvalue of Q and $C_0 > 0$ is a constant. Hence for $c \in (0, \lambda_{min}/4)$

$$\mathbb{E}[e^{c|\mathbf{x}_0|^2}] \leq e^{C_0} \int_{\mathbb{R}^d} e^{(c - \lambda_{min}/4)|x|^2} dx = e^{C_0} (2\pi)^{d/2} (\lambda_{min}/2 - 2c)^{-d/2} < \infty.$$

Thus (23) follows.

Step (ii). We will show that

$$(24) \quad p_{1-t}^*(x) \leq C \exp\left(\frac{m^2}{\sigma^2} |x|^2\right) p_t(x).$$

To this end, use the change-of-variable formula to get

$$e^{\frac{d}{2} \int_t^1 \beta_r dr} \int_{\mathbb{R}^d} \frac{\phi(y)}{p_1(y)} q(0, y, 1 - t, x) dy = \int_{\mathbb{R}^d} \frac{\phi(\sigma_{t,1} z + m_{t,1} x)}{p_1(\sigma_{t,1} z + m_{t,1} x)} \frac{e^{-|z|^2/2}}{(2\pi)^{d/2}} dz.$$

Put $w = \sigma_{t,1}z + m_{t,1}x$. Then

$$\begin{aligned}\frac{p_1(w)}{\phi(w)} &= \sigma^{-d} \exp((1/2)(1 - 1/\sigma^2)|w|^2) \int_{\mathbb{R}^d} \exp\left((m/\sigma^2)w^\top x' - (m^2/(2\sigma^2))|x'|^2\right) \mu_{data}(dx') \\ &\geq \sigma^{-d} \exp(-(m^2/(2\sigma^2))|w|^2) \mathbb{E}[e^{-(m/(\sigma^2))|\mathbf{x}_0|^2}].\end{aligned}$$

Put $\eta = \sqrt{2\mathbb{E}|\mathbf{x}_0|^2}$. By Chebyshev's inequality,

$$\begin{aligned}\mathbb{E}[e^{-(m/(\sigma^2))|\mathbf{x}_0|^2}] &\geq \mathbb{E}[e^{-(m/(\sigma^2))|\mathbf{x}_0|^2} 1_{\{|\mathbf{x}_0| < \eta\}}] \geq e^{-(m/(\sigma^2))\eta^2} \mathbb{P}(|\mathbf{x}_0| < \eta) \\ &= e^{-(m/(\sigma^2))\eta^2} (1 - \mathbb{P}(|\mathbf{x}_0| \geq \eta)) \geq e^{-(m/(\sigma^2))\eta^2} (1 - \mathbb{E}|\mathbf{x}_0|^2/(\eta^2)) \\ &= \frac{1}{2} e^{-(m/(\sigma^2))\eta^2}.\end{aligned}$$

Note that $m/(\sigma^2) \leq m = \bar{\alpha}_n$. If $\bar{\alpha}_n < 1/4$ then

$$\begin{aligned}\frac{p_{1-t}^*(x)}{p_t(x)} &\leq 2\sigma^d e^{(m/(\sigma^2))\eta^2} e^{(m^2/(\sigma^2))|x|^2} \int_{\mathbb{R}^d} e^{(m^2/(\sigma^2))|z|^2} \phi(z) dz \\ &\leq 2e^{(m/(\sigma^2))\eta^2} e^{(m^2/(\sigma^2))|x|^2} \left(1 - \frac{2m^2}{\sigma^2}\right)^{-d/2} \\ &\leq 2e^{(m/(\sigma^2))\eta^2 + 2dm^2/(\sigma^2)} e^{(m^2/(\sigma^2))|x|^2}\end{aligned}$$

where we have used $(1 - \kappa)^{-1/2} < e^\kappa$ for $0 < \kappa < 1/2$. So there exists $\delta_1 \in (0, 1/4)$ such that if $\bar{\alpha}_n < \delta_1$ then $e^{(m/(\sigma^2))\eta^2 + 2dm^2/(\sigma^2)} \leq 2$. This means

$$\frac{p_{1-t}^*(x)}{p_t(x)} \leq C e^{(m^2/(\sigma^2))|x|^2}.$$

Thus (24) follows.

Step (iii). Put $\theta(1 - t, x) = s(1 - t, x) - \nabla \log p_{1-t}(x)$. Then by Step (ii), for any $\lambda > 0$,

$$\begin{aligned}\mathbb{E}^*|\theta(1 - t_i, X_{t_i}^*)|^2 &= \int_{\{p_t^*/p_{1-t} \leq \lambda\}} |\theta(1 - t_i, x)|^2 p_t^*(x) dx + \int_{\{p_t^*/p_{1-t} > \lambda\}} |\theta(1 - t_i, x)|^2 p_t^*(x) dx \\ &\leq \lambda \mathbb{E}|\theta(1 - t_i, X_{1-t_i})|^2 + \frac{1}{\lambda} \mathbb{E}[|\theta(1 - t_i, X_{1-t_i})|^2 e^{(2m^2/(\sigma^2))|X_{1-t_i}|^2}].\end{aligned}$$

By (H2), for sufficiently small $\epsilon > 0$,

$$\begin{aligned}\mathbb{E}[e^{\epsilon|X_t|^2}] &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} e^{\epsilon|x|^2} p(0, \xi, t, x) dx p_{data}(\xi) d\xi = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} e^{\epsilon|m_{0,t}\xi + \sigma_{0,t}z|^2} \phi(z) dz p_{data}(\xi) d\xi \\ &\leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} e^{2\epsilon(|\xi|^2 + |z|^2)} \phi(z) dz p_{data}(\xi) d\xi = \mathbb{E}[e^{2\epsilon|\mathbf{x}_0|^2}] (1 - 4\epsilon)^{-d/2} \\ &\leq e^{4\epsilon d} \mathbb{E}[e^{2\epsilon|\mathbf{x}_0|^2}]\end{aligned}$$

as well as

$$\begin{aligned}
\mathbb{E}[|X_t|^2 e^{\epsilon|X_t|^2}] &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |x|^2 e^{\epsilon|x|^2} p(0, \xi, t, x) dx p_{data}(\xi) d\xi \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |m_{0,t}\xi + \sigma_{0,t}z|^2 e^{\epsilon|m_{0,t}\xi + \sigma_{0,t}z|^2} \phi(z) dz p_{data}(\xi) d\xi \\
&\leq 2 \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} (|\xi|^2 + |z|^2) e^{2\epsilon(|\xi|^2 + |z|^2)} \phi(z) dz p_{data}(\xi) d\xi \\
&= 2\mathbb{E}[|\mathbf{x}_0|^2 e^{2\epsilon|\mathbf{x}_0|^2}] (1 - 4\epsilon)^{-d/2} + 2d\mathbb{E}[e^{2\epsilon|\mathbf{x}_0|^2}] (1 - 4\epsilon)^{-d/2-1} \\
&\leq 2de^{4(d+2)\epsilon} \mathbb{E}[(1 + |\mathbf{x}_0|^2) e^{2\epsilon|\mathbf{x}_0|^2}]
\end{aligned}$$

where again we have used $(1 - \kappa)^{-1/2} < e^\kappa$ for $0 < \kappa < 1/2$. Note that $\lim_{\epsilon \rightarrow 0} \mathbb{E}[(1 + |\mathbf{x}_0|^2) e^{2\epsilon|\mathbf{x}_0|^2}] = 1 + \mathbb{E}|\mathbf{x}_0|^2$, and so there exists $\delta_2 \leq \delta_1$ such that if $\bar{\alpha}_n < \delta_2$ then

$$e^{8(d+2)(m/(\sigma^2))} \mathbb{E}[(1 + |\mathbf{x}_0|^2) e^{4(m^2/(\sigma^2))|\mathbf{x}_0|^2}] \leq 2 + \mathbb{E}|\mathbf{x}_0|^2.$$

Thus, by Lemma 2 and (H2),

$$\begin{aligned}
\mathbb{E}[|\theta(1 - t_i, X_{1-t_i})|^2 e^{(2m^2/(\sigma^2))|X_{1-t_i}|^2}] &\leq \frac{C}{m_{1-t_i}^4} \mathbb{E}[(1 + |X_{1-t_i}|^2) e^{(2m^2/(\sigma^2))|X_{1-t_i}|^2}] \\
&\leq Cd(\bar{\alpha}_n)^{-4}
\end{aligned}$$

provided that $\bar{\alpha}_n < \delta_2$.

Consequently,

$$\begin{aligned}
\sum_{i=0}^{n-1} \mathbb{E}^* |\theta(1 - t_i, X_{t_i}^*)|^2 \int_{t_i}^{t_{i+1}} \beta_{1-t} dt &\leq C \sum_{i=0}^{n-1} (-\log \alpha_{n-i}) \left(\lambda \mathbb{E} |\theta(1 - t_i, X_{1-t_i})|^2 + \frac{Cd(\bar{\alpha}_n)^{-4}}{\lambda} \right) \\
&\leq C(-n \log \min_{1 \leq i \leq n} \alpha_i) F(\lambda),
\end{aligned}$$

where

$$F(\lambda) = \lambda L + \frac{Cd(\bar{\alpha}_n)^{-4}}{\lambda}.$$

It is elementary to find that $\lambda^* = \operatorname{argmin}_{\lambda > 0} F(\lambda)$ is uniquely given by $\lambda^* = \sqrt{Cd(\bar{\alpha}_n)^{-4}/L}$, whence $\min_{\lambda > 0} F(\lambda) = C\sqrt{d(\bar{\alpha}_n)^{-4}L}$. \square

Proof of Lemma 8. Step (i). For $t < 1$ we find

$$\begin{aligned}
p_{1-t}(y) &= \int_{\mathbb{R}^d} p(0, x, 1 - t, y) p_{data}(x) dx = \int_{\mathbb{R}^d} e^{\frac{d}{2} \int_t^1 \beta_{1-r} dr} q(t, y, 1, x) p_{data}(x) dx \\
&= \mathbb{E}[p_{data}(Y_1^{t,y}) e^{\frac{d}{2} \int_t^1 \beta_{1-r} dr}].
\end{aligned}$$

Using

$$\nabla_y q(t, y, 1, z) = -f(t) \left(\frac{1}{m_{0,1-t}} y - z \right) q(t, y, 1, z),$$

we get

$$\nabla \log p_{1-t}(y) = \frac{\nabla_y \mathbb{E}[p_{data}(Y_1^{t,y})]}{\mathbb{E}[p_{data}(Y_1^{t,y})]} = -f(t) \frac{\mathbb{E}\left[\left(\frac{1}{m_{0,1-t}}y - Y_1^{t,y}\right) p_{data}(Y_1^{t,y})\right]}{\mathbb{E}[p_{data}(Y_1^{t,y})]},$$

where $f(t) = m_{0,1-t}/(\sigma_{0,1-t}^2)$. Hence, for $t < 1$,

$$\begin{aligned} \nabla \log p_{1-t}(X_t^*) &= \nabla \log p_{1-t}(Y_t) = -f(t) \frac{\mathbb{E}\left[\left(\frac{1}{m_{0,1-t}}Y_t - Y_1\right) p_{data}(Y_1) \mid \mathcal{F}_t^*\right]}{\mathbb{E}[p_{data}(Y_1) \mid \mathcal{F}_t^*]} \\ &= -f(t) \frac{\mathbb{E}\left[\left(\frac{1}{m_{0,1-t}}Y_t - Y_1\right) \frac{d\mathbb{P}^*}{d\mathbb{P}} \mid \mathcal{F}_t^*\right]}{\mathbb{E}\left[\frac{d\mathbb{P}^*}{d\mathbb{P}} \mid \mathcal{F}_t^*\right]} \\ &= -f(t) \mathbb{E}^*\left[\frac{1}{m_{0,1-t}}X_t^* - X_1^* \mid \mathcal{F}_t^*\right]. \end{aligned}$$

Applying the product Itô formula for $e^{\frac{1}{2} \int_0^{1-t} \beta_u du} X_t^*$, we derive

$$(25) \quad Y_t^* = f(t) \mathbb{E}^*\left[\int_t^1 g(r) Y_r^* dr \mid \mathcal{F}_t^*\right]$$

where $g(r) = \beta_{1-r} e^{\frac{1}{2} \int_0^{1-r} \beta_u du}$.

Step (ii). Consider the function

$$v(t, x) := \mathbb{E}^{t,x,*} \left[\int_t^1 g(r) \nabla \log p_{1-r}(X_r^*) dr \right], \quad 0 \leq t < 1,$$

where $\mathbb{E}^{t,x,*}$ is the expectation under the probability law of X^* with initial condition (t, x) . Since the transition density p^* satisfies the corresponding Kolmogorov forward equation, we have

$$-\partial_t v(t, x) = \mathcal{A}_t v(t, x) + g(t) \nabla \log p_{1-t}(x), \quad 0 \leq t < 1, \quad x \in \mathbb{R}^d,$$

where

$$\mathcal{A}_t f(x) = \left[\frac{1}{2} \beta_{1-t} x + \beta_{1-t} \nabla \log p_{1-t}(x) \right]^\top \nabla f(x) + \frac{1}{2} \beta_{1-t} \Delta f(x).$$

By the definition of v and (25) we have $Y_t^* = f(t) v(t, X_t^*)$, from which we get

$$dY_t^* = \gamma(t) Y_t + f(t) \sqrt{\beta_{1-t}} \nabla v(t, X_t^*)^\top dW_t^*,$$

where $\gamma(t) = d \log f(t) / dt - g(t) f(t) = -\beta_{1-t}/2$. Therefore with the process $Z_t^* := f(t) \sqrt{\beta_{1-t}} \nabla v(t, X_t^*)$ the representation (17) follows.

Step (iii). To estimate $|\nabla v(t, x)|$, observe

$$(26) \quad \mathbb{E}^{t,x,*} |X_r^*|^2 \leq \left(|x|^2 + 2c_0(\bar{\alpha}_n)^{-1/2} + d(-\log \bar{\alpha}_n) \right) (\bar{\alpha}_n)^{-1/2} e^{(2c_0+c_1)(\bar{\alpha}_n)^{-1}},$$

as in Lemma 4.

On the other hand, by Lemma 3,

$$\nabla_x p^*(t, x, r, y) = e^{\frac{d}{2} \int_t^r \beta_{1-u} du} p_{1-r}(y) q(t, x, r, y) (-p_{1-t}^{-2}(x)) \nabla p_{1-t}(x) = p^*(t, x, r, y) (-\nabla \log p_{1-t}(x)),$$

and so again by Lemma 2

$$\begin{aligned} |\nabla_x \mathbb{E}^{t,x,*} [\nabla \log p_{1-r}(X_r^*)]| &\leq |\nabla \log p_{1-t}(x)| \sqrt{\mathbb{E}^{t,x,*} |\nabla \log p_{1-r}(X_r^*)|^2} \\ &\leq \sqrt{2} \left(\frac{c_0}{m_{0,1-t}} + \frac{c_1}{m_{0,1-t}^2} |x| \right) \left(\frac{c_0}{m_{1-r}} + \frac{c_1}{m_{0,1-r}^2} \sqrt{\mathbb{E}^{t,x,*} |X_r^*|^2} \right). \end{aligned}$$

This and (26) yield

$$|\nabla_x \mathbb{E}^{t,x,*} [\nabla \log p_{1-r}(X_r^*)]| \leq C(\bar{\alpha}_n)^{-5/2} (\sqrt{d} + |x|^2) e^{(c_0+c_1/2)(\bar{\alpha}_n)^{-1}}$$

Thus,

$$|\nabla v(t, x)| \leq C(\bar{\alpha}_n)^{-3} (\sqrt{d} + |x|^2) e^{(c_0+c_1/2)(\bar{\alpha}_n)^{-1}} \int_0^{1-t} \beta_r dr,$$

whence

$$\begin{aligned} \mathbb{E}^* |\nabla v(t, X_t^*)|^2 &\leq C(\bar{\alpha}_n)^{-6} (d + \mathbb{E}^* |X_t^*|^4) e^{(2c_0+c_1)(\bar{\alpha}_n)^{-1}} \left(\int_0^{1-t} \beta_r dr \right)^2 \\ &\leq C(\bar{\alpha}_n)^{-6} (d + d^2 (\bar{\alpha}_n)^{-4} e^{(8c_0+6c_1)(\bar{\alpha}_n)^{-1}}) e^{(2c_0+c_1)(\bar{\alpha}_n)^{-1}} \left(\int_0^{1-t} \beta_r dr \right)^2 \\ &\leq C d^2 (\bar{\alpha}_n)^{-10} e^{(10c_0+7c_1)(\bar{\alpha}_n)^{-1}} \left(\int_0^{1-t} \beta_r dr \right)^2. \end{aligned}$$

Since $f(t) = e^{\frac{1}{2} \int_0^{1-t} \beta_r dr} / (e^{\int_0^{1-t} \beta_r dr} - 1)$ and $e^\eta - 1 \geq \eta$ for $\eta > 0$, we see

$$\int_{t_1}^{t_2} f(t)^2 \beta_{1-t} \left(\int_0^{1-t} \beta_r dr \right)^2 dt \leq \int_{t_1}^{t_2} \beta_{1-t} e^{\int_0^{1-t} \beta_r dr} dt = e^{\int_0^{1-t_1} \beta_r dr} - e^{\int_0^{1-t_2} \beta_r dr},$$

from which we obtain

$$\mathbb{E}^* \int_{t_1}^{t_2} |Z_t^*|^2 dt \leq C d^2 (\bar{\alpha}_n)^{-10} e^{(10c_0+7c_1)(\bar{\alpha}_n)^{-1}} \left(e^{\int_0^{1-t_1} \beta_r dr} - e^{\int_0^{1-t_2} \beta_r dr} \right).$$

This completes the proof of the lemma. □

Acknowledgements

This study is supported by JSPS KAKENHI Grant Number JP24K06861.

References

- [1] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P. A. Heng, and S. Z. Li. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 2024.
- [2] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan. WaveGrad: Estimating gradients for waveform generation. In *International Conference on Learning Representations*, 2021.
- [3] S. Chen, S. Chewi, J. Li, Y. Li, A. Salim, and A. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023.
- [4] R. Chetrite, P. Muratore-Ginanneschi, and K. Schwieger. E. Schrödinger’s 1931 paper “On the Reversal of the Laws of Nature” [“Über die Umkehrung der Naturgesetze”, Sitzungsberichte der preussischen Akademie der Wissenschaften, physikalisch-mathematische Klasse, 8 N9 144–153]. *Eur. Phys. J. H*, 46:1–29, 2021.
- [5] H. Chung and J. C. Ye. Score-based diffusion models for accelerated MRI. *Medical image analysis*, 80:102479, 2022.
- [6] V. De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. 2022, [arXiv:2208.05314\[stat.ML\]](#).
- [7] V. De Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.
- [8] L. C. Evans. *Partial differential equations*. American Mathematical Society, Providence, 1998.
- [9] U. G. Haussmann and E. Pardoux. Time reversal of diffusions. *Ann. Probab.*, 14:1188–1205, 1986.
- [10] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- [11] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet. Video diffusion models. 2022, [arXiv:2204.03458\[cs.CV\]](#).
- [12] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim. Diff-TTS: A denoising diffusion model for text-to-speech. In *Interspeech*, 2021.
- [13] I. Karatzas and S. E. Shreve. *Brownian motion and stochastic calculus*. Springer-Verlag, New York, 1991.
- [14] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. DiffWave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021.

- [15] H. Lee, J. Lu, and Y. Tan. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882, 2022.
- [16] H. Lee, J. Lu, and Y. Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985. PMLR, 2023.
- [17] J. S. Lee, J. Kim, and P.M. Kim. Score-based generative modeling for de novo protein design. *Nat. Comput. Sci.*, 3:382–392, 2023.
- [18] C. Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete Contin. Dyn. Syst.*, 34:1533–1574, 2013.
- [19] G. Li, Y. Wei, Y. Chen, and Y. Chi. Towards faster non-asymptotic convergence for diffusion-based generative models. 2023.
- [20] G. Li and Y. Yan. Adapting to unknown low-dimensional structures in score-based diffusion models. 2024.
- [21] H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, and Y. Chen. SRDiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.
- [22] R. Liptser and A. N. Shiryaev. *Statistics of random processes: I. General theory*. Springer, Berlin, 2nd rev. and exp. edition, 2001.
- [23] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11020–11028, 2022.
- [24] J. M. Lopez Alcaraz and N. Strodthoff. Diffusion-based time series imputation and forecasting with structured state space models. *Transactions on Machine Learning Research*, 2023.
- [25] S. Luo and W. Hu. Score-based point cloud denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4583–4592, 2021.
- [26] S. Luo, Y. Su, X. Peng, S. Wang, J. Peng, and J. Ma. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. *Advances in Neural Information Processing Systems*, 35:9754–9767, 2022.
- [27] S. D. Mbacke and O. Rivasplata. A note on the convergence of denoising diffusion probabilistic models. 2023.
- [28] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J. Y. Zhu, and S. Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.
- [29] C. Peng, P. Guo, S. K. Zhou, V. M. Patel, and R. Chellappa. Towards performant and reliable undersampled MR reconstruction via diffusion model sampling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 623–633, 2022.

- [30] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. 2022, [arXiv:2204.06125\[cs.CV\]](#).
- [31] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [32] L. Rüschemdorf and W. Thomsen. Note on the Schrödinger equation and I-projections. *Statist. Probab. Lett.*, 17:369–375, 1993.
- [33] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.
- [34] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.
- [35] Y. Song, L. Shen, L. Xing, and S. Ermon. Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations*, 2022.
- [36] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.
- [37] Y. Tashiro, J. Song, Y. Song, and S. Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems*, 34:24804–24816, 2021.
- [38] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.
- [39] T. Xie, X. Fu, O. E. Ganea, R. Barzilay, and T. S. Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. In *International Conference on Learning Representations*, 2022.
- [40] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56:1–39, 2023.
- [41] R. Yang, P. Srivastava, and S. Mandt. Diffusion probabilistic modeling for video generation. *Entropy*, 25:1469, 2023.
- [42] Q. Zhang and Y. Chen. Fast sampling of diffusion models with exponential integrator. In *International Conference on Learning Representations*, 2023.
- [43] M. Zhao, F. Bao, C. Li, and J. Zhu. EGSDE: Unpaired image-to-image translation via energy-guided stochastic differential equations. *Advances in Neural Information Processing Systems*, 35:3609–3623, 2022.