Improved Analysis of Score-based Generative Modeling: User-Friendly Bounds under Minimal Smoothness Assumptions

Hongrui Chen [*1] **Holden Lee** [*2] **Jianfeng Lu** [*3]

Abstract

We give an improved theoretical analysis of score-based generative modeling. Under a score estimate with small L^2 error (averaged across timesteps), we provide efficient convergence guarantees for any data distribution with second-order moment, by either employing early stopping or assuming a smoothness condition on the score function of the data distribution. Our result does not rely on any log-concavity or functional inequality assumption and has a logarithmic dependence on the smoothness. In particular, we show that under only a finite second moment condition, approximating the following in reverse KL divergence in ϵ-accuracy can be done in $\tilde{O}\left(\frac{d\log(1/\delta)}{\epsilon}\right)$ steps: 1) the variance-δ Gaussian perturbation of any data distribution; 2) data distributions with $1/\delta$-smooth score functions. Our analysis also provides a quantitative comparison between different discrete approximations and may guide the choice of discretization points in practice.

1. Introduction

Generative modeling is one of the central tasks in machine learning, which aims to learn a probability distribution from data and generate data from the learned distribution. Score-based generative modeling (SGM) has achieved state-of-art performance in data generation tasks (Song & Ermon, 2019; Song et al., 2020; 2021b; Dhariwal & Nichol, 2021), surpassing other models like generative adversarial networks (GAN) (Goodfellow et al., 2014), normalizing flows (Rezende & Mohamed, 2015), variational autoencoders (Kingma & Welling, 2014), and energy-based models (Zhao et al., 2016). Due to the impressive sample quality, SGM has great potential in various applications, including computer vision (Dhariwal & Nichol, 2021; Rombach et al., 2021), natural language processing (Austin et al., 2021), inverse problems (Song et al., 2022; Chung et al., 2021), molecular graph modeling (Shi et al., 2021; Gnaneshwar et al., 2022), reinforcement learning (Wang et al., 2022), and solving high-dimensional PDEs (Boffi & Vanden-Eijnden, 2022).

The key idea of SGM is to use a forward process to diffuse the data distribution to some prior (often the standard Gaussian), and learn a backward process to transform the prior to the data distribution by estimating the score functions of the forward diffusion process. Such a procedure provides an expressive and efficient way to model high-dimensional distributions for two reasons: 1) It is easy to construct a forward process that converges fast to the Gaussian, no matter how complex the data distribution is. For example, the Ornstein-Uhlenbeck (OU) process has stationary distribution equal to the standard Gaussian and converges rapidly. 2) Several scalable score matching methods such as denoising score matching (Vincent, 2011) and sliced score matching (Song et al., 2019) allow us to learn the score function for use by the backward process.

While SGM has achieved great success in practice, theoretical understanding of the power of SGM is far from complete. Recent works (Lee et al., 2022b; Chen et al., 2022) established that when an accurate score estimator is given, SGM can sample from general distributions with polynomial complexity and without requiring structural assumptions such as log-concavity or functional inequalities. (By *polynomial complexity* we mean that the running time is polynomial and the final error depends polynomially on the score estimation error and other parameters.) This is surprising in the sampling context, as it implies a sharp contrast between SGM and sampling dynamics with gradient flow structure (such as Langevin dynamics), where convergence rates depend crucially on the structure of the data distribution. In this paper, we further establish the effectiveness of SGM by showing that convergence with reasonable rates requires very weak smoothness conditions. Indeed, we obtain a logarithmic

[*]Equal contribution [1]School of Mathematical Science, Peking University [2]Applied Mathematics and Statistics Department, Johns Hopkins University [3]Department of Mathematics, Duke University. Correspondence to: Hongrui Chen <hongrui_chen@pku.edu.cn>, Holden Lee <hlee283@jhu.edu>, Jianfeng Lu <jianfeng@math.duke.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

dependence on the smoothness, or no dependence when comparing against a slightly perturbed data distribution.

1.1. Background and Our Setting

General Framework. Let P be the data distribution on $\mathbb{R}^d$. Given data $\{x_i\}_{i=1}^n$ sampled from the data distribution P, the first step of SGM involves gradually transforming the data distribution into white noise by a forward SDE:

$$\mathrm{d}\, x_t = f(x_t, t)\,\mathrm{d}\, t + g(t)\,\mathrm{d}\, w_t,\ x_0 \sim P,\ 0 \le t \le T. \quad (1)$$

We use $p_t(x)$ to denote the density of x_t. In particular, p_T is close to the white noise distribution $\mathcal{N}(0, I_d)$. Then x_t also satisfies the reverse SDE

$$\mathrm{d}\, x_t = \big(f(x_t, t) - g(t)^2 \nabla \log p_t(x_t)\big)\,\mathrm{d}\, t + g(t)\,\mathrm{d}\, \tilde{w}_t, \quad (2)$$

where $\tilde{w}_t$ is the Brownian motion in reverse time (Anderson, 1982). For convenience, we rewrite the reverse SDE (2) in a forward version by switching time direction $t \to T - t$:

$$\mathrm{d}\, \tilde{x}_t = \big(-f(\tilde{x}_t, T-t) + g(T-t)^2 \nabla \log p_{T-t}(\tilde{x}_t)\big)\,\mathrm{d}\, t + g(T-t)\,\mathrm{d}\, w_t, \quad (3)$$

where w_t is the usual (forward) Brownian motion. The process $(\tilde{x}_t)_{0 \le t \le T}$ transforms noise into samples from P, which accomplishes the goal of generative modeling.

However, we cannot directly simulate (3) since the score function $\nabla \log p_t$ is not available. Thus we learn the score function $\nabla \log p_t$ from the noisy data. First, we parameterize the score function within a function class such as that of neural networks, $s_\theta(x, t)$. Then we optimize one of the score-matching objectives (denosing score matching (Vincent, 2011) is often used; see appendix A for details), from which we obtain a score estimator s_θ such that the L^2 score estimation error

$$\mathbb{E}_{p_t} \left\| s_\theta(x, t) - \nabla \log p_t(x) \right\|^2$$

is small. Using the estimated score, we can generate samples from an approximation of the reverse SDE starting from the prior distribution:

$$\mathrm{d}\, y_t = \big(-f(y_t, T-t) + g(T-t)^2 s_\theta(y_t, T-t)\big)\,\mathrm{d}\, t + g(T-t)\,\mathrm{d}\, w_t,\ y_0 \sim p_{\text{prior}},\ 0 \le t \le T. \quad (4)$$

The Choice of Forward Process. We focus on the case $f(x, t) = -\frac{1}{2}x$, $g(t) \equiv 1$. The choice of $f(x, t)$ matches the choice in the original paper (Song et al., 2020), though our analysis may be adapted for some other choices of drift terms; the choice of constant variance function does not cause any loss of generality since the changing the variance function is equivalent to rescaling time (when f does not depend on t). In this case, the forward process becomes the Ornstein-Uhlenbeck process, which has an explicit conditional density:

$$x_t | x_0 \sim \mathcal{N}\left(e^{-\frac{1}{2}t} x_0,\ \left(1 - e^{-t}\right) I_d\right).$$

Moreover, the Ornstein-Uhlenbeck process converges exponentially to the standard Gaussian distribution:

$$\mathrm{KL}(p_t \| \mathcal{N}(0,1)) \le e^{-t} \mathrm{KL}(p_0 \| \mathcal{N}(0,1)).$$

Time Discretization. In practice, we need to use a discrete-time approximation for the sampling dynamics (4). Let $\delta = t_0 \le t_1 \le \cdots \le t_N = T$ be the discretization points, where $\delta = 0$ for the normal setting and $\delta > 0$ for the early-stopping setting. For the k-th discretization step ($1 \le k \le N$), we denote $h_k := t_k - t_{k-1}$ as the step size. We will compare different choices of discretization points and identify the optimal choice in different settings.

Let $t'_k = T - t_{N-k}$ be the corresponding discretization points in the reverse SDE. We consider two types of discretization schemes, which are widely used in existing work.

- The Euler-Maruyama scheme:

$$\mathrm{d}\hat{y}_t = \left[\frac{1}{2}\hat{y}_{t'_k} + s_\theta\left(\hat{y}_{t'_k}, T - t'_k\right)\right]\mathrm{d}t + \mathrm{d}w_t, \quad (5)$$

  for $t \in \left[t'_k, t'_{k+1}\right]$, $k = 0, 1, \ldots, N-1$.

- The expontential integrator scheme (Song et al., 2021a; Zhang & Chen, 2022): by using the semi-linear structure of (2), we discretize only in the nonlinear term and retain the continuous dynamics arising from the linear term:

$$\mathrm{d}\,\hat{y}_t = \left[\frac{1}{2}\hat{y}_t + s_\theta(\hat{y}_{t'_k}, T - t'_k)\right]\mathrm{d}\, t + \mathrm{d}\, w_t, \quad (6)$$

  for $t \in [t'_k, t'_{k+1}]$, $k = 0, \ldots, N-1$, which is solved explicitly by

$$\begin{aligned}\hat{y}_{t'_{k+1}} =& e^{\frac{1}{2}\left(t'_{k+1} - t'_k\right)} \hat{y}_{t'_k} \\ &+ 2\left(e^{\frac{1}{2}\left(t'_{k+1} - t'_k\right)} - 1\right) s_\theta(\hat{y}_{t'_k}, T - t'_k) \\ &+ \sqrt{e^{t'_{k+1} - t'_k} - 1} \cdot \eta_k,\end{aligned}$$

  where $\eta_k \sim \mathcal{N}(0, I_d)$.

1.2. Related Work

We highlight two recent papers (Chen et al., 2022; Lee et al., 2022b). Both papers provide convergence guarantees with polynomial complexity without relying on any structural assumptions on the data distribution such as log-concavity

or a functional inequality. In particular, the analysis of (Chen et al., 2022) is based on the Girsanov change of measure framework and the authors consider the following two settings: 1) The score functions in the whole trajectory of the forward process satisfy the Lipschitz condition with a uniform Lipschitz constant. 2) The data distribution has bounded support. Although the smoothness condition on the forward process seems mild, it may be hard to check whether the uniform bound for the Lipschitz constants scales polynomially w.r.t. the dimension d. In fact, this is a property of the whole process, related to tail bounds of the data distribution. The work (Lee et al., 2022b) alternatively uses the idea of excluding bad sets in order to reduce to the setting of an L^∞-accurate score estimator. This results in a worse dependence on the problem parameters; however, they do relax the smoothness condition on the whole trajectory to one on only the data distribution, and the bounded support assumption to sufficient tail decay.

Many other works have provided convergence analyses, but do not achieve polynomial complexity except in restricted settings, for example relying on functional inequalities (thus precluding multi-modal distributions) (Block et al., 2020; Lee et al., 2022a; Wibisono & Yang, 2022), manifold hypotheses (DeBortoli, 2022), or L^∞-accurate score estimates (DeBortoli et al., 2021). In the setting where only an L^2-accurate score estimate of the data distribution is given, (Koehler et al., 2022) give a statistical lower bound which shows it is in general impossible to accurately sample the distribution. This highlights the fact that having score estimates for multiple distributions—e.g., the data distribution with different amounts of noise added—is necessary for efficient sampling; this is done in practice and in our analysis. In a different direction, SGM is also related to recent work on algorithmic stochastic localization (Alaoui et al., 2022), in which for the spin glass models under consideration, the score function (i.e., the posterior mean) can be accurately estimated using approximate message passing.

1.3. Our Contributions

In this paper, we quantitatively show that an L^2-accurate score estimator is enough to guarantee that the sampling dynamics (5), (6) result in a distribution close to the data distribution in various regimes. Our results combine the advantages of (Chen et al., 2022; Lee et al., 2022b): under weak assumptions on the data distribution and the score estimator, we provide a concise analysis and refined guarantees for the convergence of SGM under several settings, described below and summarized in Table 1.

Smooth setting. Revisiting the setting where the Lipshitz constant of $\nabla \log p_t$, $0 \le t \le T$ is uniformly bounded (the trajectory-smooth setting), we provide three refinements compared to (Chen et al., 2022): 1) We sidestep the technical issue of checking Novikov's condition and provide a reverse KL divergence guarantee, which is stronger than a TV guarantee. 2) For the exponential integrator scheme, the number of steps dependends logarithmically rather than polynomially on the second moment. 3) We do not assume the data distribution has finite KL divergence wrt the standard Gaussian.

Non-smooth setting. We provide convergence guarantees for sampling from any distribution with bounded second-order moment, without any structural assumption or smoothness condition. In particular, for any small constant $\delta > 0$, we show that running the sampling dynamics (6) with appropriate early stopping and decreasing step size results in a distribution close to p_δ, using a high-probability bound on the Hessian matrix $\nabla^2 \log p_t$ and a change-of-measure argument. Comparing to the early stopping result in (Chen et al., 2022), the use of a high-probability rather than uniform bound on the Hessian removes the bounded support assumption and induces a significantly tighter dependence on the problem parameters. Quantitatively, to obtain a bound of ϵ_{TV} in TV-distance to p_δ, when the data distribution is supported on a ball of radius R, (Chen et al., 2022) require $\tilde{\Theta}\left(\frac{dR^4}{\epsilon_{\mathrm{TV}}^2 \delta^4}\right)$ steps, while we consider a distribution with second moment bounded by M_2 and only require $\tilde{\Theta}\left(\frac{d^2}{\epsilon_{\mathrm{TV}}^2} \log^2 \frac{M_2 d}{\delta}\right)$ steps (typically, $R \asymp \sqrt{d}$). We have no dependence on R, and our dependence on δ and M_2 is logarithmic instead of polynomial.

By adding an extra truncation step on the algorithm, we also obtain a pure Wasserstein bound depending on the tail decay of the data distribution, significantly improving the prior result (Lee et al., 2022a, Theorem 2.2).

Smooth p_0 only. Finally, we consider the intermediate assumption of smoothness of $\nabla \log p_0$, rather than the whole forward process as in (Chen et al., 2022). In this case, we can bound discretization error in the low-noise regime so that early stopping is not required. We combine the smooth and non-smooth analyses to bound the number of steps logarithmically in L, the Lipschitz constant of $\nabla \log p_0$.

Furthermore, we analyze difference choices of discretization schemes and step-size schedules (equivalently, different variance functions). This may help guide the practical implementation of SGM.

1.4. Notations

General Notations. Let d be the dimension of the data, and γ_d be the density of standard Gaussian measure $\mathcal{N}(0, I_d)$. $\|\cdot\|$ denotes the ℓ^2 norm for vectors or the spectral norm for matrices, and $\|\cdot\|_F$ denotes the Frobenius norm of matrices. For a random variable X, the sub-exponential

Table 1. Suppose p_0 has bounded 2nd moment M_2 and average L^2 score error is at most ϵ_0^2. Guarantees for DDPM hold under the following smoothness assumptions, listed in order of decreasing strength. Note the 2nd bound also holds under the 1st assumption, but trades off dependence on d and L. (Chen et al., 2022) obtain TV guarantees, which are weaker by Pinsker's inequality.

ASSUMPTION	ERROR GUARANTEE	STEPS TO GET $\tilde{O}(\epsilon_0^2)$ ERROR	THEOREM
$\forall t$, $\nabla \log p_t$ L-LIPSCHITZ	$\mathrm{KL}(p_0 \Vert \hat{q}_T)$	$\tilde{O}\left(\frac{dL^2}{\epsilon_0^2}\right)$	THEOREM 2.1
	$\mathrm{TV}(p_0, \hat{q}_T)^2$	$\tilde{O}\left(\frac{(d \vee M_2)L^2}{\epsilon_0^2}\right)$	(CHEN ET AL., 2022, THEOREM 2)
$\nabla \log p_0$ L-LIPSCHITZ	$\mathrm{KL}(p_0 \Vert \hat{q}_T)$	$\tilde{O}\left(\frac{d^2 \log^2 L}{\epsilon_0^2}\right)$	THEOREM 2.5
NONE	$\mathrm{KL}(p_\delta \Vert \hat{q}_{T-\delta})$	$\tilde{O}\left(\frac{d^2 \log^2(1/\delta)}{\epsilon_0^2}\right)$	THEOREM 2.2
SUPPORTED ON $B_R(0)$	$\mathrm{TV}(p_\delta, \hat{q}_{T-\delta})^2$	$\tilde{O}\left(\frac{(d \vee M_2)R^4}{\epsilon_0^2 \delta^4}\right)$	(CHEN ET AL., 2022, THM. 2 + LEM. 16)

and sub-gaussian norms are defined by

$$\|X\|_{\psi_k} := \inf\{t > 0 : \mathbb{E} \exp\left(|X|^k / t\right) \le 2\},\ k = 1, 2.$$

For random vectors, we denote $\|\cdot\|_{\psi_k} := \||\cdot|\|_{\psi_k}$. We use $x \asymp y$ if there exist absolute constants $C_1, C_2 > 0$ such that $C_1 y \le x \le C_2 y$. Write $x \lesssim y$ to mean $x \le Cy$ for an absolute constant $C > 0$, and define $x \gtrsim y$ analogously.

Notations for the Forward Process. Let P be the data distribution and p_0 be its density (if it exists). For $0 < t \le T$, let p_t be the density of x_t defined in the forward process (1) with $f(t, x) = \frac{1}{2} g(t)^2 x_t$. Define σ_t as the conditional variance of x_t given x_0, i.e.,

$$\sigma_t^2 := 1 - e^{-t}.$$

For any $0 \le t \le s \le T$, let

$$\alpha_{t,s} := e^{-\frac{1}{2}(s-t)}, \quad \alpha_t := \alpha_{0,t}$$

gives the scaling between times t and s: $\mathbb{E}[x_s | x_t] = \alpha_{t,s} x_t$.

Notations for Reverse Processes. Let $s(x, t)$ be the estimated score function. The reverse processes arising in our setting are defined as follows:

- Let $\tilde{x}_t$ be the the reverse process of $(x_t)_{0 \le t \le T}$, which is driven by the SDE

$$\mathrm{d}\,\tilde{x}_t = \left(\frac{1}{2}\tilde{x}_t + \nabla \log p_{T-t}(\tilde{x}_t)\right)\mathrm{d}\,t + \mathrm{d}\,w_t,\ \tilde{x}_0 \sim p_T$$

  Then the law of $(\tilde{x}_t)_{0 \le t \le T}$ is identical to the law of $(x_{T-t})_{0 \le t \le T}$. We use $\tilde{p}_t$ to denote the density of $\tilde{x}_t$.

- Let $\hat{y}_t$ be the discrete approximation of y_t defined in (5) or (6) starting from $\hat{y}_0 \sim \mathcal{N}(0, I_d)$. We use $\hat{q}_t$ to denote the density of $\hat{y}_t$.

2. Main Results

We first consider the trajectory smoothness assumption, where we strengthen the result of (Chen et al., 2022). Then, we state our results for more general settings in various regimes.

All the results rely on L^2-accuracy of the score estimator:

Assumption 1. The learned score function $s(x, t)$ satisfies for any $1 \le k \le N$,

$$\frac{1}{T}\sum_{k=1}^{N} h_k\, \mathbb{E}_{p_{t_k}} \|\nabla \log p_{t_k}(x) - s(x, t_k)\|^2 \le \epsilon_0^2. \tag{7}$$

Remark 1. Because this is a weighted average of score estimation errors on the discretization points, it can be satisfied even if the error diverges as $t \to 0$. This is useful because simply based on the size of the gradient, we can expect the error to scale as $\mathbb{E}_{p_{t_k}} \|\nabla \log p_{t_k}(x) - s(x, t_k)\|^2 \lesssim \frac{\epsilon^2}{\sigma_{t_k}^2}$, where $\sigma_t^2 \sim t$ as $t \to 0$. The calculation $\int_{t_1}^{1} \frac{1}{t}\, dt = \log(1/t_1)$ tells us we can take $\epsilon_0^2 = O(\epsilon^2 \log(1/t_1))$. See Appendix A for details.

Assumption 2. The data distribution has a bounded second moment: $M_2 := \mathbb{E}_P \|x\|^2 < \infty$.

2.1. Analysis under the Trajectory Smoothness Condition

First, we improve result of (Chen et al., 2022) for the trajectory-smooth setting, weakening the assumptions and strengthening the conclusion.

Assumption 3. For any $0 \le t \le T$, $\nabla \log p_t$ is L-Lipschitz on $\mathbb{R}^d$.

Theorem 2.1. *Suppose that Assumptions 1,2,3 hold. If $L \ge 1$, $h_k \le 1$ for $k = 1, \ldots, N$ and $T \ge 1$, using uniform discretization points yields the followings*

- *Using exponential integrator scheme (6), we have*

$$\mathrm{KL}(p_0 \Vert \hat{q}_T) \lesssim (M_2 + d)e^{-T} + T\epsilon_0^2 + \frac{dT^2L^2}{N}.$$

In particular, choosing $T = \log\left(\frac{M_2+d}{\epsilon_0^2}\right)$ *and* $N = \Theta\left(\frac{dT^2L^2}{\epsilon_0^2}\right)$ *makes this* $\widetilde{O}\left(\epsilon_0^2\right)$.

- *Using the Euler-Maruyama scheme (5), we have*

$$\mathrm{KL}(p_0\|\hat{q}_T) \lesssim (M_2+d)e^{-T} + T\epsilon_0^2 + \frac{dT^2L^2}{N} + \frac{T^3M_2}{N^2}.$$

For the exponential integrator, the error consists of three parts: the error of the forward process, the score matching error, and the discretization error, detailed in Section 3.

Remark 2. The extra conditions on L, h_k, T in the above theorem are introduced to present the result more concisely, and are not a limitation of the analysis.

Remark 3. Comparing to the exponential integrator scheme, the Euler-Maruyama scheme causes an additional high-order discretization error term related to the second-order moment of the data distribution. This implies a separation between the exponential integrator scheme and the Euler-Maruyama scheme: the error of the exponential integrator scheme scales logarithmically in the second moment of the data distribution (as it suffices for T to increase by $O(\log M_2)$), while the error of the Euler-Maruyama scheme scales linearly.

Rather than the TV distance guarantees given in (Chen et al., 2022), we obtain (reverse) KL divergence guarantees which are stronger by Pinsker's inequality and nontrivial even when the bound is larger than 1.

Discussion for Lipschitzness Assumption 3. Though Assumption 3 seems mild, it is hard to check whether the Lipschitz constant of the score function is bounded uniformly by a constant $L = O(\mathrm{poly}(d))$ throughout the entire process. In the log-concave setting, the smoothness of $\nabla \log p_0$ implies the smoothness of $\nabla \log p_t$ (Lee et al., 2021, Lemma 28). However, for non-log-concave distributions such as multi-modal distributions, this can be difficult to check, and may depend on the tail behavior of the data distribution. Our aim in this work is to relax such smoothness assumptions.

2.2. Results for General Distributions with Early Stopping

We now consider the most general setting: we provide convergence guarantees for any distribution that has a bounded second-order moment, without introducing any structural assumptions or smoothness conditions. Hence, our results are applicable to the case that the score function is non-smooth or even not well defined, like distributions supported on a low-dimensional manifold.

Due to our weak assumptions, the backward process (2) may have very bad properties when t is close to 0, so we need to employ early stopping. For any small constant $\delta > 0$, we show that running the sampling dynamics (6) for time $T - \delta$ will result in a distribution close to p_δ in KL divergence. Note that in general, it is impossible to obtain KL or TV closeness to P as this requires matching exactly the support of P.

We provide the convergence bound for general discretization and further quantify the bound for several specific choices.

Theorem 2.2. *There is a universal constant* K *such that the following hold. Suppose that Assumptions 1 and 2 hold and the step sizes satisfy*

$$\frac{h_k}{\sigma^2_{t_{k-1}}} \le \frac{1}{Kd}, \quad k = 1, \dots, N. \tag{8}$$

Define $\Pi := \sum_{k=1}^N \frac{h_k^2}{\sigma^4_{t_{k-1}}}$. *For* $T \ge 2, \delta \le \frac{1}{2}$, *the exponential integrator scheme (6) with early stopping result in a distribution* $\hat{q}_{T-\delta}$ *such that*

$$\mathrm{KL}(p_\delta\|\hat{q}_{T-\delta}) \lesssim (d+M_2)\exp(-T) + T\epsilon_0^2 + d^2\Pi. \tag{9}$$

In particular, for exponentially decreasing step size $h_k = c\min\{t_k, 1\}$, *where* $c \le \frac{1}{Kd}$ *(or, equivalently* $\frac{\log\left(\frac{1}{\delta}\right)+T}{N} \le \frac{1}{Kd}$*), then (8) holds and*

$$\Pi \lesssim \frac{\left(\log\left(\frac{1}{\delta}\right)+T\right)^2}{N}.$$

Choosing $T = \log\left(\frac{M_2+d}{\epsilon_0^2}\right), N = \Theta\left(\frac{\left(\log\left(\frac{1}{\delta}\right)+T\right)^2 d^2}{\epsilon_0^2}\right)$ *makes this* $\widetilde{O}\left(\epsilon_0^2\right)$.

In addition, for the Euler-Maruyama scheme (5), the same bounds hold with an additional term $M_2\sum_{k=1}^N h_k^3$ *term in the right hand side of (9).*

Remark 4. Note the upper bound (9) works for any choice of discretization points. We will quantify the term Π for other choices of discretization in the later paragraph.

Remark 5. By rescaling time, choosing constant variance function $g \equiv 1$ and exponentially decreasing step size is equivalent to choosing exponential g and constant step size. We state the theorem with constant g for convenience (with an exponential choice of g, we would only reach the data distribution P at time $t = -\infty$).

The key difficulty in analyzing general distributions is that the discretization error is hard to control without the Lipschitz condition on $\nabla \log p_t$. Our approach is to use a high-probability bound for the Hessian matrix $\nabla^2 \log p_t$ with a change of measure. This approach works well for constant-order t, while in the low-noise regime the bound will explode as t tends to 0. We overcome the blow-up of discretization error by early stopping.

Discussion on the Choice of Discretization Points. When t goes to 0, the regularity of $\nabla \log p_t$ becomes worse so slowing down the SDE leads to a smaller discretization error. In the result of Theorem 2.2, the term $\Pi = \sum_{k=1}^{N} \frac{h_k^2}{\sigma^4_{t_{k-1}}}$ in the upper bound (9) depends on the choice of discretization points. In particular,

- If we choose uniform discretization $h_k = c$, the dependence on $\frac{1}{\delta}$ becomes linear.
- (Song et al., 2020) considers variance function $g(t) = \sqrt{t}$ with uniform discreitzation. This is equivalent to using constant variance function with quadratic discretization points $t_k = (\delta + kh)^2$ for appropriate h. This choice of discretization points induces a linear step size and our Theorem results in a square-root dependence on $\frac{1}{\delta}$.
- In Theorem 2.2, by using exponentially decaying (and then constant) step size, we reduce this error to a logarithmic dependence. Indeed, the term Π achieves its minimum (up to a constant) under our choice of discretization points.

See Appendix B for details. Although under our assumptions, the theory suggests that exponentially decreasing step sizes are optimal, other issues may arise in practice. We leave an experimental comparison of different g's or step sizes to future work.

Wasserstein+KL Guarantee. Notice that when δ is small, p_δ is only a small perturbation (in Wasserstein distance) of the data distribution P. Then stopping the algorithm at appropriate δ results in a distribution that is close in KL divergence to a distribution that is close to P in Wasserstein distance, and we obtain the following.

Corollary 2.3. *Suppose that Assumptions 1 and 2 hold for data distribution P. Then using the exponential integrator scheme with exponentially decreasing step size, to reach a distribution Q such that $W_2^2(P, M_\sharp p_\delta) \le \epsilon_{\mathrm{W}}^2 \le \frac{d}{2}$ and $\mathrm{KL}(M_\sharp p_\delta \| Q) \le \epsilon_{\mathrm{KL}}^2 \le \frac{d+M_2}{2}$ requires*

$$N = \Theta\left(\frac{d^2 \log^2\left(\frac{(d+M_2)d}{\epsilon_{\mathrm{KL}}^2 \epsilon_{\mathrm{W}}^2}\right)}{\epsilon_{\mathrm{KL}}^2}\right)$$

steps and Assumption 1 to hold with

$$\epsilon_0^2 \le \frac{\epsilon_{\mathrm{KL}}^2}{K \log^2\left(\frac{d+M_2}{\epsilon_{\mathrm{KL}}^2}\right)}$$

for an appropriate absolute constant K. Here, $M(x) = \exp(\frac{\delta}{2})x$, $M_\#$ denote the pushforward map on the distributions, $Q = M_\sharp \hat{q}_{T-\delta}$.

Remark 6. Corollary 2.3 implies an upper bound for the bounded Lipschitz metric between the data distribution and $\hat{p}_{t_0}$(as mentioned in (Chen et al., 2022)):

$$\sup\left\{\mathbb{E}_\mu f - \mathbb{E}_\nu f : \mid f : \mathbb{R}^d \to [-1,1] \text{ is 1-Lipschitz}\right\}.$$

Note our improved dependencies compared with (Chen et al., 2022, Corollary 3) and (Lee et al., 2022b, Theorem 2.1).

While the smoothness assumption is relaxed, our analysis induces an additional d-factor in place of the Lipschitz constant of $\nabla \log p_t$ compared to Theorem 2.1. This d-factor comes from the high-probability bound for the Hessian matrix (see Lemma C.7). However, (Chen et al., 2022, Theorem 5) suggests that the lower bound of the discretization error scales linearly on d. We leave open the problem of closing the gap between the dimension dependence in the upper and lower bounds.

Pure Wasserstein Guarantee. We can also obtain a pure Wasserstein guarantee by following (Lee et al., 2022a, Theorem 2.2). For this, we need to include an extra truncation step on the algorithm output, i.e., for some choice of R, replacing any sample $\hat{y}_{T-\delta} \sim \hat{q}_{T-\delta}$ falling outside $B_R(0)$ by 0. In addition, we need to assume some concentration for P, so that samples from P lie in $B_R(0)$ with high probability.

Corollary 2.4. *Consider the distribution $\hat{q}^{\mathrm{trunc}}_{T-\delta}$ obtained by exponential integrator scheme with exponentially decreasing step size and the truncation step. Suppose that Assumptions 1 and 2 hold with $\epsilon_0 = O\left(\frac{\epsilon_{\mathrm{W}}^2}{R^2}\right)$, and that $R \ge M_2, \delta, T, N$ satisfy*

$$\begin{aligned} &\delta = \Theta\left(\frac{\epsilon_{\mathrm{W}}^2}{d}\right), \quad R^2 \mathbb{P}(\|x_\delta\| \ge R) = O(\epsilon_{\mathrm{W}}^2), \\ &T = \Theta\left(\log\left(\frac{R^4(M_2+d)}{\epsilon_{\mathrm{W}}^4}\right)\right), \\ &N = \Theta\left(\frac{d^2 R^4 \left(\log\frac{1}{\delta} + T\right)^2}{\epsilon_{\mathrm{W}}^4}\right). \end{aligned} \tag{10}$$

Then the resulting truncated and scaled distribution $M_\sharp \hat{q}^{\mathrm{trunc}}_{T-\delta}$ satisfies $W_2^2(P, M_\sharp \hat{q}^{\mathrm{trunc}}_{T-\delta}) = \tilde{O}(\epsilon_{\mathrm{W}}^2)$. (Here, M is as in Corollary 2.3.)

Remark 7. Note that the appropriate R in (10) exists under mild tail conditions on the data distribution P. For example:

- If there exists a constant $\eta > 0$ such that $\mathbb{E}_P \|x\|^{2+\eta} = O(\mathrm{poly}(d))$, R depends polynomially on $\frac{1}{\epsilon_{\mathrm{W}}}$ and d and thus we obtain a polynomial complexity guarantee.
- When the data distribution P is K sub-exponential, R has a logarithmic dependence on $\frac{1}{\epsilon_{\mathrm{W}}}$ and (10) induces $N = \tilde{O}\left(\frac{d^2 K^4}{\epsilon_{\mathrm{W}}^4}\right)$.

2.3. Result for Smooth Data Distributions

We further provide convergence analysis for smooth p_0 without using early stopping. As mentioned in Subsection 2.2, the early stopping technique is employed to bound the discretization error in the low-noise regime. We can alternatively bound this error by using the smoothness condition on p_0:

Assumption 4. The data distribution admits a density $p_0 \in C^2(\mathbb{R}^d)$ and $\nabla \log p_0$ is L-Lipschitz.

We bound the discretization error in two different time regimes: Choosing an appropriate constant $\delta_0 > 0$, when $t > \delta_0$, we use a high-probability Hessian bound and a change of measure argument similar to the analysis in the early stopping setting; for $t < \delta_0$, we alternatively derive a Lipschitz constant bound for $\nabla \log p_t$ (stated in Lemma C.9) based on Assumption 4.

Theorem 2.5. *There is a universal constant K such that the following holds. Under Assumptions 1, 2, and 4, by using the exponentially decreasing (then constant) step size $h_k = c\min\{\max\{t_k, \frac{1}{L}\}, 1\}$, $c = \frac{\log L + T}{N} \le \frac{1}{Kd}$, the sampling dynamic (6) results in a distribution $\hat{q}_T$ such that*

$$\mathrm{KL}(p_0 \| \hat{q}_T) \lesssim (M_2 + d)\exp(-T) + T\epsilon_0^2 + \frac{d^2(\log L + T)^2}{N}.$$

Choosing $T = \log\left(\frac{M_2+d}{\epsilon_0^2}\right)$ and $N = \Theta\left(\frac{d^2(T+\log L)^2}{\epsilon_0^2}\right)$ makes this $\widetilde{O}\left(\epsilon_0^2\right)$.

In addition, for Euler-Maruyama scheme (5), the same bounds hold with an additional $M_2 \sum_{k=1}^N h_k^3$ term.

Comparing to Theorem 2.1, this result only depends on the Lipschitz constant of $\nabla \log p_0$ rather than the uniform Lipschitz constant bound for $\nabla \log p_t$, $0 \le t \le T$. We also ease the dependency on L from L^2 to $\log^2 L$ for optimal choice of variance function or step size, so the requirement on the smoothness of the data distribution is significantly relaxed: even if the Lipschitz constant L scales exponentially on d, we can still obtain a polynomial complexity guarantee. Note that we do pay an extra d factor compared to Theorem 2.1.

3. Proof sketches

We sketch the proofs of the main theorems using the exponential integrator discretization, and give complete proofs in Appendices C and D. We first consider the smooth setting, and then describe the modifications for the non-smooth case. Our main technical novelty lies in the arguments for the non-smooth setting, we also streamline the arguments in the smooth setting and use an interpolation rather than Girsanov approach that gives KL divergence bounds.

3.1. Smooth setting (Theorem 2.1)

First term. The first source of error arises from the mismatch between the distribution of the forward process p_T at time T, and our Gaussian initialization for the reverse process, $\hat{q}_0 = \gamma_d$. We can separate out this term using the chain rule for KL divergence:

$$\mathrm{KL}(p_0 \| \hat{q}_T) \le \mathrm{KL}(p_T \| \hat{q}_0) + \mathbb{E}_{p_T(a)}\, \mathrm{KL}(p_{0|T}(\cdot|a) \| \hat{q}_{T|0}(\cdot|a)).$$

The first term can be bounded using exponential mixing of the forward (Ornstein-Uhlenbeck) process towards the standard Gaussian. In conjunction with the fact that after constant time, the KL-divergence is bounded by $O(d + M_2)$, we obtain (Lemma C.4)

$$\mathrm{KL}(p_T \| \hat{q}_0) \lesssim (d + M_2)e^{-T}.$$

Note this estimate does not depend on the initial distance $\mathrm{KL}(p_0 \| \gamma_d)$ as in (Chen et al., 2022).

The remaining term can be written as a sum, again using the chain rule for KL divergence, by comparing the continuous process with the estimated, discrete process through a chain of intermediate processes where we run the continuous process until time t_k. We can interpolate the discrete processes to realize them as SDE's. If Novikov's conditions are satisfied, Girsanov's Theorem then applies to bound the KL divergence in terms of the squared difference of the drift terms between the processes.

$$\begin{aligned}
&\mathbb{E}_{p_T(a)}\, \mathrm{KL}(p_{T|0}(\cdot|a) \| \hat{q}_{T|0}(\cdot|a)) \\
&= \sum_{k=1}^N \mathbb{E}_{p_{t_k}(a)}\, \mathrm{KL}(p_{t_{k-1}|t_k}(\cdot|a) \| \hat{q}_{T-t_{k-1}|T-t_k}(\cdot|a)) \\
&\le \sum_{k=1}^N \frac{1}{2} \int_{t_{k-1}}^{t_k} \mathbb{E}_{x_t \sim p_t} \left\| s(x_{t_k}, t_k) - \nabla \log p_t(x_t) \right\|^2 dt \\
&\le \underbrace{\sum_{k=1}^N \int_{t_{k-1}}^{t_k} \mathbb{E}_{x_t \sim p_t} \left\| s(x_{t_k}, t_k) - \nabla \log p_{t_k}(x_{t_k}) \right\|^2}_{(2)} \\
&\qquad + \underbrace{\sum_{k=1}^N \mathbb{E} \left\| \nabla \log p_{t_k}(x_{t_k}) - \nabla \log p_t(x_t) \right\|^2 dt}_{(3)}.
\end{aligned}$$

In the last step we use the triangle inequality. However, in general Novikov's condition may not be satisfied; (Chen et al., 2022) circumvent this using an involved truncation argument which only results in a TV bound and relies on the trajectory-smooth condition (Assumption 3). We instead use a differential inequality argument which gives the same conclusion (Lemma C.1, C.2, Proposition C.3) and is applicable to the non-smooth setting; this step requires significant technical work (Appendix F).

Second term. Term (2) is exactly the score estimation error, and by Assumption 1, it is bounded by $T\epsilon_0^2$.

Third term. Term (3) is the discretization error. This discretization error bound is non-trivial since in classical numerical analysis theory, the discretization error often depends exponentially on the time T due to the use of Gronwall's inequality. Our analysis our will rely on the special structure of the Ornstein-Uhlenbeck process. We note that (3) involves both a "time" and "space" discretization error (as both the time and space arguments are different). We show in Lemma C.6 that this can be bounded purely in terms of the space discretization error (which streamlines the argument of (Chen et al., 2022))

$$\mathbb{E}\left\|\nabla \log p_s(x_s) - \nabla \log p_t(x_t)\right\|^2 \lesssim (s-t)^2 \cdot \mathbb{E}\left\|\nabla \log p_t(x_t)\right\|^2 + \mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_t(\alpha_{s,t}^{-1} x_s)\right\|^2 .$$

The explicit form of the OU process tells us that $\alpha_{s,t}^{-1} x_s = x_t + z$, where z is a Gaussian of variance $O(s-t)$. Therefore, the second term (which dominates) can be bounded as a Lipschitz constant times the second moment of a Gaussian:

$$\mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_t(\alpha_{s,t}^{-1} x_s)\right\|^2 \lesssim L^2\, \mathbb{E}\left\|z\right\|^2 \lesssim dL^2(s-t). \quad (11)$$

Note that we crucially use the Lipschitzness of the score in this step. Plugging this bound into the sum (3) gives the final error term.

3.2. Non-smooth setting (Theorem 2.2)

Comparing Theorem 2.1 (smooth setting) and Theorem 2.2 (non-smooth setting), we note that the discretization error changes from $\frac{T^2 L^2 d}{N}$ to $\frac{\left(\log\left(\frac{1}{\delta}\right)+T\right)^2 d}{N}$; the intuition is that L is "effectively" bounded by $\sqrt{d}$. Previously, (Chen et al., 2022) assume that P is supported on a ball of radius R to derive a global Lipschitzness bound $\left\|\nabla^2 \log p_t\right\| = O\left(\frac{R^2}{t^2}\right)$ to plug into the smooth theorem.

Our main insight is that (1) because we are averaging the error over p_t, it suffices to have a high-probability rather than uniform bound on the the Hessian, and (2) such bounds are obtainable from the smoothing properties of the forward process. In fact, to bound (11), we only need Lipschitzness in a *random* direction, and hence a Frobenius norm bound is sufficient (Lemma C.7):

$$\left\|\left\|\nabla^2 \log p_t(x)\right\|_F\right\|_{\psi_1} \lesssim \frac{d}{\min\{t,1\}}. \quad (12)$$

(This is the weaker analogue of an operator norm bound of $O(\sqrt{d})$, which was suggested from the $L = O(\sqrt{d})$ analogy.) This incurs significant savings over a uniform bound, and in particular does not depend on boundedness or tails of P. We prove this by giving a Bayesian interpretation of the Hessian as the posterior variance of the noise in the score matching objective. As a purely mathematical statement about smoothing of the OU process, this result may be of independent interest.

Finally, to use (12) in (11), we actually need to bound the Hessian not just at x_t but along the path (in direction z) joining x_t and $\alpha_{s,t}^{-1} x_s$: for this we need a change-of-measure argument (Lemma C.8) which says that the distributions of (x_t, z) and $(x_t + az, z)$ are close in χ^2-divergence, for $0 \le a \le 1$. Finally, although the bound (12) blows up as $t \to 0$, by choosing an exponentially decreasing step size and stopping at time δ, we only incur a $\log\left(\frac{1}{\delta}\right)$ dependence, similarly to the analysis of the score estimation error (Remark 1).

3.3. Smooth p_0 (Theorem 2.5)

If we only assume $\nabla \log p_0$ is L-Lipschitz, we can still derive Lipschitzness of $\nabla \log p_0$ for small time $t \le \frac{1}{L}$ (Lemma C.9). For large $t \ge L$, the argument in the non-smooth case applies (and gives a bound of $O(dL)$ in (12)). Thus, we take exponentially decreasing step size until $t = 1/L$, and then constant step size, and combine the analyses of Theorems 2.1 and 2.2 to obtain Theorem 2.5.

4. Conclusion

In this paper, we analyzed the theoretical properties of SGM in various regimes. We extended existing result to the most general setting and provided refined guarantees. The current analysis provides guarantees for SGM in the framework that an L^2-accurate score estimator is available. This implies the training objective in denoising score matching is suitable for learning a generative model and partially explains why SGM is empirically successful at modeling very complex distributions, like multi-mode distributions or distributions with weak smoothness condition.

We obtain guarantees for arbitrary data distributions without smoothness assumptions, by exploiting (high-probability) smoothing properties of the forward process. Besides closing the factor-d gap between our upper bound and the (suggested) lower bound, it would be interesting to carry out this kind of analysis for other choices of the forward/backward processes, such as critically damped Langevin Diffusion (Dockhorn et al., 2021), to see if improved guarantees are available. ((Chen et al., 2022) show that no improvement is available only in the setting of a uniform bound on the Lipschitz constant of the score.)

Another future direction is to explore theories beyond the framework that an L^2-accurate score estimator is available and understand the learning of a score estimator, includ-

ing the approximability, sample complexity, and the training dynamics of denoising score matching. This is related to the most challenging problems in deep learning theory; advances in deep learning theory may provide some new insight into SGM.

Acknowledgement

The work of JL is supported in part by National Science Foundation through award DMS-2012286. HC is partially supported by the elite undergraduate training program of School of Mathematical Sciences in Peking University.

References

Alaoui, A. E., Montanari, A., and Sellke, M. Sampling from the sherrington-kirkpatrick gibbs measure via algorithmic stochastic localization. *arXiv preprint arXiv:2203.05093*, 2022.

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

Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and van den Berg, R. Structured denoising diffusion models in discrete state-spaces. In *NeurIPS*, 2021.

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

Boffi, N. M. and Vanden-Eijnden, E. Probability flow solution of the fokker-planck equation, 2022. arXiv:2206.04642.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. R. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions, 2022. URL https://arxiv.org/abs/2209.11215.

Chewi, S., Erdogdu, M. A., Li, M., Shen, R., and Zhang, S. Analysis of langevin monte carlo from poincare to log-sobolev. In Loh, P.-L. and Raginsky, M. (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 1–2. PMLR, 02–05 Jul 2022.

Chung, H., Sim, B., and Ye, J.-C. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction, 2021. arXiv:2112.05146.

DeBortoli, V. Convergence of denoising diffusion models under the manifold hypothesis, 2022. arXiv:2208.05314.

DeBortoli, V., Thornton, J., Heng, J., and Doucet, A. Diffusion schrödinger bridge with applications to score-based generative modeling. In *NeurIPS*, 2021.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 2021.

Dockhorn, T., Vahdat, A., and Kreis, K. Score-based generative modeling with critically-damped langevin diffusion. *arXiv preprint arXiv:2112.07068*, 2021.

Gnaneshwar, D., Ramsundar, B., Gandhi, D., Kurchin, R. C., and Viswanathan, V. Score-based generative models for molecule generation, 2022. arXiv:2203.04698.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. In *NIPS*, 2014.

Karatzas, I. and Shreve, S. *Brownian motion and stochastic calculus*, volume 113. Springer Science & Business Media, 1991.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes, 2014. arXiv:1312.6114.

Koehler, F., Heckett, A., and Risteski, A. Statistical efficiency of score matching: The view from isoperimetry, 2022. URL https://arxiv.org/abs/2210.00726.

Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28:1302–1338, 2000.

Lee, H., Pabbaraju, C., Sevekari, A., and Risteski, A. Universal approximation for log-concave distributions using well-conditioned normalizing flows. *arXiv preprint arXiv:2107.02951*, 2021.

Lee, H., Lu, J., and Tan, Y. Convergence for score-based generative modeling with polynomial complexity, 2022a. URL https://arxiv.org/abs/2206.06227.

Lee, H., Lu, J., and Tan, Y. Convergence of score-based generative modeling for general data distributions, 2022b. URL https://arxiv.org/abs/2209.12381.

Rezende, D. J. and Mohamed, S. Variational inference with normalizing flows. In *ICML*, 2015.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. *CVPR*, 2021.

Shi, C., Luo, S., Xu, M., and Tang, J. Learning gradient fields for molecular conformation generation. In *ICML*, 2021.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models, 2021a. arXiv:2010.02502.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Song, Y., Garg, S., Shi, J., and Ermon, S. Sliced score matching: A scalable approach to density and score estimation. In *UAI*, 2019.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

Song, Y., Durkan, C., Murray, I., and Ermon, S. Maximum likelihood training of score-based diffusion models. In *NeurIPS*, 2021b.

Song, Y., Shen, L., Xing, L., and Ermon, S. Solving inverse problems in medical imaging with score-based generative models, 2022. arXiv:2111.08005.

Vempala, S. S. and Wibisono, A. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In *NeurIPS*, 2019.

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

Wang, Z., Hunt, J. J., and Zhou, M. Diffusion policies as an expressive policy class for offline reinforcement learning, 2022. arXiv:2208.06193.

Wibisono, A. and Yang, K. Y. Convergence in kl divergence of the inexact langevin algorithm with application to score-based generative models, 2022. URL https://arxiv.org/abs/2211.01512.

Zhang, Q. and Chen, Y. Fast sampling of diffusion models with exponential integrator, 2022. arXiv:2204.13902.

Zhao, J. J., Mathieu, M., and LeCun, Y. Energy-based generative adversarial network, 2016. arXiv:1609.03126.

A. Denoising Score Matching

For $0 \le t \le T$, the goal of score matching for p_t is to minimize

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

Since the score function $\nabla \log p_t$ is not available, we alternatively consider a denoising score matching objective (Vincent, 2011), which is derived from integrating by parts

$$\begin{aligned}
&\mathbb{E}_{p_t} \|s_\theta(x,t) - \nabla \log p_t(x)\|^2 \\
&= \mathbb{E}_{p_t} \|s_\theta(x,t)\|^2 + \mathbb{E}_{p_t} \|\nabla \log p_t\|^2 - 2\,\mathbb{E}_{p_t} \langle s_\theta(x,t), \nabla \log p_t(x)\rangle \\
&= \mathbb{E}_{p_t} \|s_\theta(x,t)\|^2 + \mathbb{E}_{p_t} \|\nabla \log p_t(x)\|^2 + 2\,\mathbb{E}_{p_t} \nabla \cdot s_\theta(x,t) \\
&= \mathbb{E}_{p_t} \|s_\theta(x,t)\|^2 + \mathbb{E}_{p_t} \|\nabla \log p_t(x)\|^2 + 2\,\mathbb{E}_{p_0(x_0)}\, \mathbb{E}_{p_{t|0}(x_t|x_0)}\, \nabla \cdot s_\theta(x_t,t) \\
&= \mathbb{E}_{p_t} \|s_\theta(x,t)\|^2 + \mathbb{E}_{p_t} \|\nabla \log p_t(x)\|^2 - 2\,\mathbb{E}_{p_0(x_0)}\, \mathbb{E}_{p_{t|0}(x_t|x_0)} \langle \nabla \log p_{t|0}(x_t|x_0), s_\theta(x_t,t)\rangle \\
&= \mathbb{E}_{p_t} \|s_\theta(x,t)\|^2 + \mathbb{E}_{p_t} \|\nabla \log p_t(x)\|^2 - 2\,\mathbb{E}_{p_0(x_0)}\, \mathbb{E}_{p_{t|0}(x_t|x_0)} \left\langle \frac{x_t - \alpha_t x_0}{\sigma_t^2}, s_\theta(x_t,t)\right\rangle \\
&= \mathbb{E} \left\| s_\theta(x_t,t) - \frac{x_t - \alpha_t x_0}{\sigma_t^2}\right\|^2 + \mathbb{E}_{p_t} \|\nabla \log p_t(x)\|^2 - \frac{d}{\sigma_t^2} \\
&= \mathbb{E} \left\| s_\theta(x_t,t) - \frac{x_t - \alpha_t x_0}{\sigma_t^2}\right\|^2 + C,
\end{aligned}$$

where $p_{t|0}$ is the conditional distribution of x_t given x_0, and C is a constant independent of θ.

Noticing that $\mathbb{E} \left\| \frac{x_t - \alpha_t x_0}{\sigma_t^2}\right\|^2 = \frac{d}{\sigma_t^2}$, it is natural to expect the error to scale as

$$\mathbb{E}_{p_{t_k}} \|\nabla \log p_{t_k}(x) - s(x,t_k)\|^2 \lesssim \frac{\epsilon^2}{\sigma_{t_k}^2}.$$

In this case, by noting that $\sigma_{t_k}^2 \asymp \min\{1, t_k\}$, we have

$$\mathbb{E}_{p_{t_k}} \|\nabla \log p_{t_k}(x) - s\left(x,t_k\right)\|^2 \lesssim \frac{\epsilon^2}{\min\left\{t_k, 1\right\}},$$

then (7) is satisfied with a log factor:

$$\frac{1}{T}\sum_{k=1}^T h_k\, \mathbb{E}_{p_{t_k}} \|\nabla \log p_{t_k}(x) - s\left(x,t_k\right)\|^2 \lesssim \frac{1}{T}\int_{t_1}^T \frac{\epsilon^2}{t \wedge 1} dt \lesssim \epsilon^2 \log\left(\frac{1}{t_1}\right)$$

B. Discussion on Choices of Discretization Points

In this section, we consider the scaling of the term $\Pi = \sum_{k=1}^N \frac{h_k^2}{\sigma_{t_{k-1}}^4}$ in (9) under different choices of discretization points.

The Constant Step Size For uniform discretization(inducing constant step size) $t_k = \delta + kh,\ h = \frac{T-\delta}{N}$, we have

$$\begin{aligned}
\sum_{k=1}^N \frac{h_k^2}{\sigma_{t_{k-1}}^4} &\asymp \sum_{t_k \le 1} \frac{h_k^2}{t_{k-1}^2} + \sum_{t_k > 1} h_k^2 \\
&\asymp h \int_\delta^1 \frac{1}{t^2}\,\mathrm{d}\,t + \frac{T^2}{N} \\
&\asymp \frac{T/\delta + T^2}{N}.
\end{aligned}$$

Thus the upper bound for discretization error has a linear dependence on $\frac{1}{\delta}$.

The Linear Step Size For quadratic discretization points(inducing linear step size) $t_k = (\delta + kh)^2$, $h = \frac{\sqrt{T}-\delta}{N}$, by noting that $\frac{h_k}{h} \asymp \sqrt{t_k}$, we have

$$\begin{aligned}\sum_{k=1}^{N} \frac{h_k^2}{\sigma^4_{t_{k-1}}} &\asymp \sum_{t_k \le 1} \frac{h_k^2}{t_{k-1}^2} + \sum_{t_k > 1} h_k^2 \\ &\asymp h \sum_{t_k \le 1} \frac{h_k}{t_k^{3/2}} + h \sum_{t_k > 1} \sqrt{t_k} h_k \\ &\asymp h \int_{\delta}^{1} \frac{1}{t_k} \,\mathrm{d}\, t + h \int_{1}^{T} \sqrt{t} \,\mathrm{d}\, t \\ &\asymp \frac{1}{N} \left(\sqrt{\frac{T}{\delta}} + T^2 \right).\end{aligned}$$

Optimality of Exponential Decaying Step Size Now we will show that the discretization points used in Theorem 2.2 minimizes the term Π (up to a constant). Indeed, note that

$$\Pi \asymp \Pi_1 + \Pi_2, \quad \Pi_1 := \sum_{t_k \le 1} \frac{(t_k - t_{k-1})^2}{t_k^2}, \, \Pi_2 := \sum_{t_k > 1} (t_k - t_{k-1})^2.$$

For the term Π_1, let $z_k = \log \frac{t_k}{t_{k-1}} > 0$, we have $\Pi = \sum_{k=1}^{n} (e^{z_k} - 1)^2$. Note that $z \mapsto (e^z - 1)^2$ is convex for $z > 0$. By Jensen's inequality, when the summation of z_k's are fixed, the minimum of Π_1 is reached when z_k's are identical. Equivalently, $h_k = ct_k$ for $t_k \le 1$. For the term Π_2, we have $\Pi_2 = \sum_{t_k > 1} h_k^2$. Similarly, since $h \mapsto h^2$ is convex for $h > 0$, the minimum of Π_2 is reached when h_k's are identical.

C. Main Proof Ingredients

The key idea of the proof is motivated by the Girsanov change of measure framework used in (Chen et al., 2022). However, in order to avoid the technical challenge of altering the process to satisfy Novikov's condition, we use a differential inequality-based argument instead.

Lemma C.1. *Consider the following two Itô processes*

$$\begin{aligned}\mathrm{d}\, X_t &= F_1(X_t, t) \,\mathrm{d}\, t + g(t) \,\mathrm{d}\, w_t, & X_0 &= a, \\ \mathrm{d}\, Y_t &= F_2(Y_t, t) \,\mathrm{d}\, t + g(t) \,\mathrm{d}\, w_t, & Y_0 &= a,\end{aligned}$$

where F_1, F_2, g are continuous functions and may depend on a. We assume the uniqueness and regularity condition:

- *The two SDEs have unique solutions.*
- *X_t, Y_t admit densities $p_t, q_t \in C^2(\mathbb{R}^d)$ for $t > 0$.*

Define the relative Fisher information between p_t and q_t by

$$J(p_t \| q_t) = \int p_t(x) \left\| \nabla \log \frac{p_t(x)}{q_t(x)} \right\|^2 \mathrm{d}\, x.$$

Then for any $t > 0$, the evolution of $\mathrm{KL}(p_t \| q_t)$ is given by

$$\frac{\partial}{\partial t} \mathrm{KL}(p_t \| q_t) = -g(t)^2 J(p_t \| q_t) + \mathbb{E}\left[\left\langle F_1(X_t, t) - F_2(X_t, t), \nabla \log \frac{p_t(X_t)}{q_t(X_t)} \right\rangle \right].$$

Remark 8. While we have written the same Brownian motion for X and Y, as we only care about distributions, the Brownian motions can be chosen independent with each other.

We will apply Lemma C.1 on $(\tilde{x}_t)_{0\le t\le T-\delta}$ and $(\hat{y}_t)_{0\le t\le T-\delta}$ to show the convergence in KL divergence. The following lemma collects some technical properties of the two processes. The proof of both lemmas is deferred to Appendix F.

Lemma C.2. *For $0 \le k \le N-1$, consider the reverse SDE starting from $\tilde{x}_{t'_k} = a$*

$$\mathrm{d}\,\tilde{x}_t = \left[\frac{1}{2}\tilde{x}_t + \nabla \log \tilde{p}_t(\tilde{x}_t)\right]\mathrm{d}\,t + \mathrm{d}\,w_t, \quad \tilde{x}_{t'_k} = a \tag{13}$$

and its discrete approximation:

$$\mathrm{d}\,\hat{y}_t = \left[\frac{1}{2}\hat{y}_t + s(a, T-t'_k)\right]\mathrm{d}\,t + \mathrm{d}\,w_t, \quad \hat{y}_{t'_k} = a \tag{14}$$

for time $t \in (t'_k, t'_{k+1}]$. Let $\tilde{p}_{t|t'_k}$ be the density of $\tilde{x}_t$ given $\tilde{x}_{t'_k}$ and $\hat{q}_{t|t'_k}$ be density of $\hat{y}_t$ given $\hat{y}_{t'_k}$. Then we have

1. *For any $a \in \mathbb{R}^d$, the two processes satisfy the uniqueness and regularity condition stated in Lemma C.1, that is, (13) and (14) have unique solution and $\tilde{p}_{t|t'_k}(\cdot|a), \hat{q}_{t|t'_k}(\cdot|a) \in C^2(\mathbb{R}^d)$ for $t > t'_k$.*
2. *For a.e. $a \in \mathbb{R}^d$ (with respect to the Lebesgue measure), we have*

$$\lim_{t\to t'_k+} \mathrm{KL}(\tilde{p}_{t|t'_k}(\cdot|a)\|\hat{q}_{t|t'_k}(\cdot|a)) = 0.$$

In addition, the above results also hold if we replace $\hat{y}_t$ with that corresponding to the Euler-Maruyama scheme:

$$\mathrm{d}\,\hat{y}_t = \left[\frac{1}{2}g(T-t)^2 a + g(T-t)^2 s_\theta(a, T-t'_k)\right]\mathrm{d}\,t + \mathrm{d}\,w_t,\ \hat{y}_{t'_k} = a.$$

Proposition C.3. *Under Assumption 1, we have*

- *The exponential integrator scheme (6) satisfies*

$$\mathrm{KL}(p_\delta\|\hat{q}_{T-\delta}) \lesssim \mathrm{KL}(p_T\|\gamma_d) + T\epsilon_0^2 + \sum_{k=1}^{N}\int_{t_{k-1}}^{t_k} \mathbb{E}\,\|\nabla \log p_t(x_t) - \nabla \log p_{t_k}(x_{t_k})\|^2\,\mathrm{d}\,t.$$

- *The Euler-Maruyama scheme (5) satisfies*

$$\begin{aligned}\mathrm{KL}(p_\delta\|\hat{q}_{T-\delta}) \lesssim\ & \mathrm{KL}(p_T\|\gamma_d) + T\epsilon_0^2 \\ & + \sum_{k=1}^{N}\int_{t_{k-1}}^{t_k} \left(\mathbb{E}\,\|\nabla \log p_t(x_t) - \nabla \log p_{t_k}(x_{t_k})\|^2 + \mathbb{E}\,\|x_t - x_{t_k}\|^2\right)\mathrm{d}\,t.\end{aligned}$$

Proof. Let us consider first the exponential integrator. For $t'_k < t \le t'_{k+1}$, let $\tilde{p}_{t|t'_k}$ be the distribution of $\tilde{x}_t$ given $\tilde{x}_{t'_k}$ and $\hat{q}_{t|t'_k}$ be the distribution of $\hat{y}_t$ given $\tilde{y}_{t'_k}$. From Lemma C.2(1) the uniqueness and regularity condition in Lemma C.1 hold for (13) and (14). Thus for any $a \in \mathbb{R}^d$ and $t > t'_k$ we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\,t}\mathrm{KL}(\tilde{p}_{t|t'_k}(\cdot|a)\|\hat{q}_{t|t'_k}(\cdot|a)) =& -\frac{1}{2}\,\mathbb{E}_{\tilde{p}_{t|t'_k}(y|a)}\left\|\nabla \log \frac{\tilde{p}_{t|t'_k}(y|a)}{\hat{q}_{t|t'_k}(y|a)}\right\|^2 \\ & + \mathbb{E}_{\tilde{p}_{t|t'_k}(y|a)}\left[\left\langle (\nabla \log \tilde{p}_t(y) - s(a, t_{N-k})), \nabla \log \frac{\tilde{p}_{t|t'_k}(y|a)}{\hat{q}_{t|t'_k}(y|a)}\right\rangle\right] \\ \le & \frac{1}{2}\,\mathbb{E}_{\tilde{p}_{t|t'_k}(y|a)}\,\|s(a, t_{N-k}) - \nabla \log \tilde{p}_t(y)\|^2,\end{aligned} \tag{15}$$

where we use the fact that $\langle v, w\rangle \le \frac{1}{2}\|v\|^2 + \frac{1}{2}\|w\|^2$. By Lemma C.2(2), for a.e. $a \in \mathbb{R}^d$ we have

$$\lim_{t\to t'_k+} \mathrm{KL}(\tilde{p}_{t|t'_k}(\cdot|a)\|\hat{q}_{t|t'_k}(\cdot|a)) = 0,$$

and hence

$$\mathrm{KL}(\tilde{p}_{t'_{k+1}|t'_k}(\cdot|a)\|\hat{q}_{t'_{k+1}|t'_k}(\cdot|a)) \le \frac{1}{2}\int_{t'_k}^{t'_{k+1}} \mathbb{E}_{\tilde{p}_{t|t'_k}(y|a)} \|s(a,t_{N-k}) - \nabla \log \tilde{p}_t(y)\|^2 \,\mathrm{d}\, t.$$

Since $\tilde{p}_{t'_k}$ is absolutely continuous w.r.t. the Lebesgue measure, integrating on the both sides w.r.t. $\tilde{p}_{t'_k}$ yields

$$\mathbb{E}_{\tilde{p}_{t'_k}(a)} \mathrm{KL}(\tilde{p}_{t'_{k+1}|t'_k}(\cdot|a)\|\hat{q}_{t'_{k+1}|t'_k}(\cdot|a)) \le \frac{1}{2}\int_{t'_k}^{t'_{k+1}} \mathbb{E} \|s(\tilde{x}_{t'_k},t_{N-k}) - \nabla \log \tilde{p}_t(\tilde{x}_t)\|^2 \,\mathrm{d}\, t.$$

For $0 \le k \le N-1$, we use the chain rule of KL divergence to obtain

$$\begin{aligned}
\mathrm{KL}(\tilde{p}_{t'_{k+1}}\|\hat{q}_{t'_{k+1}}) &\le \mathbb{E}_{\tilde{p}_{t'_k}(a)} \mathrm{KL}(\tilde{p}_{t'_{k+1}|t'_k}(\cdot|a)\|\hat{q}_{t'_{k+1}|t'_k}(\cdot|a)) + \mathrm{KL}(\tilde{p}_{t'_k}\|\hat{q}_{t'_k}) \\
&\le \mathrm{KL}(\tilde{p}_{t'_k}\|\hat{q}_{t'_k}) + \frac{1}{2}\int_{t'_k}^{t'_{k+1}} \mathbb{E} \|s(\tilde{x}_{t'_k},T-t'_k) - \nabla \log \tilde{p}_t(\tilde{x}_t)\|^2 \,\mathrm{d}\, t.
\end{aligned}$$

Summing over $k = 0, 1, \ldots, N-1$ and using $p_t = \tilde{p}_{T-t}$, we obtain

$$\begin{aligned}
\mathrm{KL}(p_\delta\|\hat{q}_{T-\delta}) &\le \mathrm{KL}(p_T\|\gamma_d) + \frac{1}{2}\sum_{k=0}^{N-1}\int_{t'_k}^{t'_{k+1}} \mathbb{E} \|s(\tilde{x}_{t'_k},T-t'_k) - \nabla \log \tilde{p}_t(\tilde{x}_t)\|^2 \,\mathrm{d}\, t \\
&\le \mathrm{KL}(p_T\|\gamma_d) + \frac{1}{2}\sum_{k=1}^{N}\int_{t_{k-1}}^{t_k} \|s(x_{t_k},t_k) - \nabla \log p_t(x_t)\|^2 \,\mathrm{d}\, t \\
&\le \mathrm{KL}(p_T\|\gamma_d) + \sum_{k=1}^{N}\int_{t_{k-1}}^{t_k} \|s(x_{t_k},t_k) - \nabla \log p_{t_k}(x_{t_k})\|^2 \,\mathrm{d}\, t \\
&\quad + \sum_{k=1}^{N}\int_{t_{k-1}}^{t_k} \|\nabla \log p_{t_k}(x_{t_k}) - \nabla \log p_t(x_t)\|^2 \,\mathrm{d}\, t \\
&\le \mathrm{KL}(p_T\|\gamma_d) + T\epsilon_0^2 + \sum_{k=1}^{N}\int_{t_{k-1}}^{t_k} \|\nabla \log p_{t_k}(x_{t_k}) - \nabla \log p_t(x_t)\|^2 \,\mathrm{d}\, t.
\end{aligned}$$

This completes the proof for the exponential integrator scheme. The proof for the Euler-Maruyama scheme is similar; the only difference is the differential inequality becomes

$$\frac{\mathrm{d}}{\mathrm{d}\, t}\mathbb{E}_{\tilde{p}_{t'_k}} \mathrm{KL}(\tilde{p}_{t|t'_k}(\cdot|x)\|\hat{q}_{t|t'_k}(\cdot|x)) \le \frac{1}{2}\mathbb{E}\left\|\nabla \log \tilde{p}_t(\tilde{x}_t) - s(\tilde{x}_{t'_k},t_{N-k}) + \frac{1}{2}(\tilde{x}_t - \tilde{x}_{t'_k})\right\|^2$$

and we can obtain the result in an analogous way. □

The three terms in the upper bound of Proposition C.3 match the claim in Theorem 2.1. The first term is controlled by the exponential convergence of the forward process, which is given in the following lemma.

Lemma C.4. *Under Assumption 2, for $T > 1$, we have*

$$\mathrm{KL}(p_T\|\gamma_d) \le (d + M_2)e^{-T}.$$

Proof. Notice that $x \mapsto x\log x$ is a convex function for $x > 0$. Let $p_{t|0}$ be the conditional density of x_t given x_0. For any $t > 0$, we can use Jensen's inequality to bound the entropy of p_t:

$$\begin{aligned}
\int_{\mathbb{R}^d} p_t(x)\log p_t(x)\,\mathrm{d}\, x &= \int_{\mathbb{R}^d}\left[\left(\int_{\mathbb{R}^d} p_{t|0}(x|y)\,\mathrm{d}\, P(y)\right)\log\left(\int_{\mathbb{R}^d} p_{t|0}(x|y)\,\mathrm{d}\, P(y)\right)\right]\mathrm{d}\, x \\
&\le \int_{\mathbb{R}^d}\left[\int_{\mathbb{R}^d} p_{t|0}(x|y)\log p_{t|0}(x|y)\,\mathrm{d}\, P(y)\right]\mathrm{d}\, x
\end{aligned}$$

$$= \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} p_{t|0}(x|y) \log p_{t|0}(x|y) \, \mathrm{d}\, x \right) \mathrm{d}\, P(y).$$

Since $x_t | x_0 = y \sim \mathcal{N}(\alpha_t x_0, \sigma_t^2 I_d)$, we have

$$\int_{\mathbb{R}^d} p_{t|0}(x|y) \log p_{t|0}(x|y) \, \mathrm{d}\, x = -\frac{d}{2} \log(2\pi\sigma_t^2) - \frac{d}{2}.$$

Thus

$$\int_{\mathbb{R}^d} p_t(x) \log p_t(x) \, \mathrm{d}\, x \le -\frac{d}{2} \log(2\pi\sigma_t^2) - \frac{d}{2}.$$

Therefore,

$$\begin{aligned} \mathrm{KL}(p_t \| \gamma_d) &= \int_{\mathbb{R}^d} p_t(x) \log p_t(x) \, \mathrm{d}\, x + \mathbb{E}_{p_t} \left[\frac{\|x\|^2}{2} + \frac{d}{2} \log(2\pi) \right] \\ &\le \frac{d}{2} \log \sigma_t^{-2} + \frac{1}{2}(M_2 - d). \end{aligned}$$

From the exponential convergence of Langevin dynamics with strongly log-concave stationary distribution (see, e.g., (Vempala & Wibisono, 2019)), we obtain

$$\mathrm{KL}(p_T \| \gamma_d) \le e^{-T+t} \left(\frac{d}{2} \log \sigma_t^{-2} + \frac{1}{2}(M_2 - d) \right).$$

By choosing $t = \log 2$, we have

$$e^t \log \left(\frac{1}{\sigma_t^2} \right) \lesssim 1.$$

Thus

$$\mathrm{KL}(p_T \| \gamma_d) \lesssim e^{-T}(d + M_2). \qquad \square$$

The second term in the upper bound of Proposition C.3 is exactly the same as the score estimation error defined in Assumption 1. So the key challenge is to bound the third term, which is caused by the discretization error.

According to Proposition C.3, the discretization error of the Euler-Maruyama scheme induces an extra linear term $\mathbb{E} \|x_t - x_{t_k}\|^2$ compared to the exponential integrator scheme. The following lemma bounds this extra term.

Lemma C.5. *Suppose that $h_k \le 1$ for $1 \le k \le N$. We have*

$$\mathbb{E} \|x_t - x_{t_k}\|^2 \lesssim d(t_k - t) + M_2(t_k - t)^2, \quad t_{k-1} \le t \le t_k,$$

and

$$\sum_{k=1}^N \int_{t_{k-1}}^{t_k} \mathbb{E} \|x_t - x_{t_k}\|^2 \, \mathrm{d}\, t \lesssim d \sum_{k=1}^N h_k^2 + M_2 \sum_{k=1}^N h_k^3.$$

Proof. From the definition of the forward process (1), we have

$$\begin{aligned} \mathbb{E} \|x_t - x_{t_k}\|^2 &= \mathbb{E} \left\| \int_t^{t_k} \frac{1}{2} x_u \, \mathrm{d}\, u - \int_t^{t_k'} \mathrm{d}\, w_u \right\|^2 \\ &\lesssim \mathbb{E} \left\| \int_t^{t_k} x_u \, \mathrm{d}\, u \right\|^2 + \left\| \int_t^{t_k} \mathrm{d}\, w_u \right\|^2 \\ &\le (t_k - t) \left(\int_t^{t_k} \mathbb{E} \|x_u\|^2 \, \mathrm{d}\, u \right) + d(t_k - t), \end{aligned} \tag{16}$$

where the last inequality follows from the Cauchy-Schwartz inequality. From the explicit form of the conditional density

$$x_u|x_0 \sim \mathcal{N}\left(e^{-\frac{1}{2}u}x_0, \left(1-e^{-u}\right) I_d\right),$$

the second moment of x_u is bounded by $\mathbb{E}\left\|x_u\right\|^2 \leq M_2 + d$. Pluging this into (16), we arrive at

$$\mathbb{E}\left\|x_t - x_{t_k}\right\|^2 \lesssim d(t_k - t) + (d + M_2)(t_k - t)^2.$$

Therefore,

$$\int_{t_{k-1}}^{t_k} \mathbb{E}\left\|x_t - x_{t_k}\right\|^2 \lesssim dh_k^2 + (d + M_2)h_k^3.$$

Taking summation over $k = 1, \ldots, N$, we complete the proof. □

Therefore, we only need to focus on the term $\mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_{t_k}(x_{t_k})\right\|^2$. This discretization is taken both in space and time. One observation is that the time-discretization error can be absorbed by the space-discretization error.

Lemma C.6. *For any $0 \leq t \leq s \leq T$, the forward process (1) satisfies*

$$\mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_s(x_s)\right\|^2 \;\leq\; 4\,\mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_t(\alpha_{t,s}^{-1}x_s)\right\|^2 + 2\,\mathbb{E}\left\|\nabla \log p_t(x_t)\right\|^2 \left(1 - \alpha_{t,s}^{-1}\right)^2.$$

Proof. Since $x_s|x_t \sim \mathcal{N}\left(\alpha_{t,s}x_t, (1-\alpha_{t,s}^2)I_d\right)$, from Lemma E.1, we can rewrite $\nabla \log p_s$ as

$$\nabla \log p_s(x) = \alpha_{t,s}^{-1}\,\mathbb{E}_{p_{t|s}(y|x)}\,\nabla_y \log p_t(y),$$

where $p_{t|s}$ is the conditional density of x_t given x_s. Thus the time discretization error can be bounded by

$$\begin{aligned}
\mathbb{E}\left\|\nabla \log p_t(\alpha_{t,s}^{-1}x_s) - \nabla \log p_s(x_s)\right\|^2 &= \mathbb{E}_{p_s}\left\|\alpha_{t,s}^{-1}\,\mathbb{E}_{p_{t|s}(y|x_s)}\,\nabla \log p_t(y) - \nabla \log p_t(\alpha_{t,s}^{-1}x_s)\right\|^2 \\
&\leq \mathbb{E}\left\|\alpha_{t,s}^{-1}\nabla \log p_t(x_t) - \nabla \log p_t(\alpha_{t,s}^{-1}x_s)\right\|^2 \\
&\leq 2(1-\alpha_{t,s}^{-1})^2\,\mathbb{E}\left\|\nabla \log p_t(x_t)\right\|^2 + 2\,\mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_t(\alpha_{t,s}^{-1}x_s)\right\|^2.
\end{aligned}$$

Therefore, splitting the error into the space-discretization and the time-discretization error,

$$\begin{aligned}
&\mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_s(\alpha_{t,s}^{-1}x_s)\right\|^2 \\
&\leq 2\,\mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_t(\alpha_{t,s}^{-1}x_s)\right\|^2 + 2\,\mathbb{E}\left\|\nabla \log p_t(\alpha_{t,s}^{-1}x_s) - \nabla \log p_s(x_s)\right\|^2 \\
&\leq 2(1-\alpha_{t,s}^{-1})^2\,\mathbb{E}\left\|\nabla \log p_t(x_t)\right\|^2 + 4\,\mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_t(\alpha_{t,s}^{-1}x_s)\right\|^2.
\end{aligned}$$

We complete the proof. □

In Lemma C.6, $(1-\alpha_{t,s}^{-1})^2 = O((s-t)^2)$ and the term $\mathbb{E}\left\|\nabla \log p_t(x_t)\right\|^2$ can be bounded by Lemma E.2, so the space-discretization error dominates the right hand side. In what follows, we tackle the space-discretization term $\mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_t(\alpha_{t,s}^{-1}x_s)\right\|^2$ in various regimes. In particular:

- If the score functions of the forward process is smooth, i.e., Assumption 3 holds, the space-discretization error can be directly bounded using the Lipschitz condition on $\nabla \log p_t$.
- In the general setting, we choose a early stopping time t_0 and bound the space-discretization error for $t > t_0$ by a high-probability bound on the Hessian matrix $\nabla^2 \log p_t$ and a change of measure argument, which are worked out in section C.1.
- For smooth p_0, we further bound the space-discretization error for small t by providing a Lipschitz constant bound for $\nabla \log p_t$ when t is sufficient small, which is given in section C.2.

C.1. The High-probability Hessian Bound and Change of Measure

In this subsection, we establish the high-probability bound for the Hessian matrix $\nabla^2 \log p_t$ and use the high-probability bound to control the space-discretization error. This is the critical part of our analysis that allows us to prove Theorem 2.2.

Lemma C.7. *Let P be a probability measure on $\mathbb{R}^d$. Consider the density its Gaussian perturbation $p_\sigma(x) \propto \int_{\mathbb{R}^d} \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \mathrm{d}\, P(y)$. Then for $x \sim p_\sigma$, we have the sub-exponential norm bound*

$$\|\nabla^2 \log p_\sigma(x)\|_{F,\psi_1} \lesssim \frac{d}{\sigma^2},$$

where $\|\cdot\|_{F,\psi_1} = \|\|\cdot\|_F\|_{\psi_1}$ denote the sub-exponential norm of the Frobenius norm of a random matrix.

Proof. Define the conditional density $\tilde{P}_\sigma(y|x)$ as $\mathrm{d}\, \tilde{P}_\sigma(y|x) \propto \exp\left(-\frac{\|y-x\|^2}{2\sigma^2}\right) \mathrm{d}\, P(y)$. Using Lemma E.3, $\nabla^2 \log p_\sigma$ can be written as

$$\nabla^2 \log p_\sigma(x) = \mathrm{Var}_{\tilde{P}_\sigma(y|x)}\left(\frac{y}{\sigma^2}\right) - \frac{I_d}{\sigma^2}.$$

For any positive integer p, using the fact that $\frac{y-x}{\sigma}$ is distributed as $\mathcal{N}(0, I_d)$ and the power mean inequality,

$$\begin{aligned}
\mathbb{E}_{p_\sigma(x)} \left\|\mathrm{Var}_{\tilde{P}_\sigma(y|x)}\left(\frac{y}{\sigma^2}\right)\right\|_F^p &\le \frac{1}{\sigma^{2p}} \mathbb{E}_{p_\sigma(x)} \left\|\mathbb{E}_{\tilde{P}_\sigma(y|x)}\left(\frac{y-x}{\sigma}\right)\left(\frac{y-x}{\sigma}\right)^\top\right\|_F^p \\
&\le \frac{1}{\sigma^{2p}} \mathbb{E}_{z\sim\mathcal{N}(0,I_d)} \|zz^\top\|_F^p. \\
&\lesssim \left(\frac{pd}{\sigma^2}\right)^p.
\end{aligned}$$

Using the arbitrariness of p, we know that

$$\left\|\mathrm{Var}_{\tilde{P}_\sigma(y|x)}\left(\frac{y}{\sigma^2}\right)\right\|_{F,\psi_1} \lesssim \frac{d}{\sigma^2}.$$

Thus by the triangle inequality,

$$\|\nabla^2 \log p_\sigma(x)\|_{F,\psi_1} \lesssim \frac{d}{\sigma^2}.$$

We complete the proof. □

Lemma C.8. *There is a universal constant $K > 0$ so that the following holds. For $0 \le t \le s \le T$, $\frac{s-t}{\sigma_t^2} \le \frac{1}{Kd}$, we have*

$$\mathbb{E}\,\|\nabla \log p_t(x_t) - \nabla \log p_t(\alpha_{t,s}^{-1} x_s)\|^2 \lesssim \frac{d^2(s-t)}{\sigma_t^4}.$$

Proof. We bound the difference between the value of $\nabla \log p_t$ at different points with the Hessian:

$$\nabla \log p_t(x_t) - \nabla \log p_t(\alpha_{t,s}^{-1} x_s) = \int_0^1 \nabla^2 \log p_t(x_t + a(\alpha_{t,s}^{-1} x_s - x_t))(\alpha_{t,s}^{-1} x_s - x_t)\, \mathrm{d}\, a.$$

Thus

$$\mathbb{E}\,\|\nabla \log p_t(x_t) - \nabla \log p_t(\alpha_{t,s}^{-1} x_s)\|^2 \le \int_0^1 \mathbb{E}\left\|\nabla^2 \log p_t(x_t + a z_{t,s}) z_{t,s}\right\|^2 \mathrm{d}\, a, \tag{17}$$

where $z_{t,s}$ is defined by $z_{t,s} = \alpha_{t,s}^{-1} x_s - x_t \sim \mathcal{N}\left(0, (e^{s-t} - 1) I_d\right)$ and is independent of x_t. For random vectors X, Y, we use $P_{X,Y}$ to denote the joint probability measure of (X, Y) and $P_{X|Y}$ to denote the conditional probability measure of X

given Y. Then for $0 \le a \le 1$, we use change of measure to bound $\mathbb{E}\left\|\nabla^2 \log p_t(x_t + a z_{t,s}) z_{t,s}\right\|^2$:

$$\begin{aligned}\mathbb{E}\left\|\nabla^2 \log p_t(x_t + a z_{t,s}) z_{t,s}\right\|^2 &= \mathbb{E}\left[\left\|\nabla^2 \log p_t(x_t) z_{t,s}\right\|^2 \frac{\mathrm{d}\, P_{x_t + a z_{t,s}, z_{t,s}}(x_t, z_{t,s})}{\mathrm{d}\, P_{x_t, z_{t,s}}(x_t, z_{t,s})}\right] \\ &\lesssim \left(\mathbb{E}\left\|\nabla^2 \log p_t(x_t) z_{t,s}\right\|^4 \mathbb{E}\left(\frac{\mathrm{d}\, P_{x_t + a z_{t,s}, z_{t,s}}(x_t, z_{t,s})}{\mathrm{d}\, P_{x_t, z_{t,s}}(x_t, z_{t,s})}\right)^2\right)^{1/2}.\end{aligned} \tag{18}$$

Let $M_t = \nabla^2 \log p_t(x_t)(\nabla^2 \log p_t(x_t))^\top$, $Z_{t,s} = z_{t,s} z_{t,s}^\top$. For $A, B \in \mathbb{R}^{d\times d}$, define the tensor product $A \otimes B \in (\mathbb{R}^d)^{\otimes 4}$ as $(A \otimes B)_{i_1,i_2,i_3,i_4} = A_{i_1 i_2} B_{i_3 i_4}$. Since M_t and $Z_{t,s}$ are independent, the first factor in (18) can be written as

$$\begin{aligned}\mathbb{E}\left\|\nabla^2 \log p_t(x_t) z_{t,s}\right\|^4 &= \mathbb{E}\left[\mathrm{Tr}\left(M_t^\top Z_{t,s}\right)^2\right] \\ &= \mathbb{E}\langle M_t \otimes M_t, Z_{t,s} \otimes Z_{t,s}\rangle \\ &= \langle \mathbb{E}\, M_t \otimes M_t, \mathbb{E}\, Z_{t,s} \otimes Z_{t,s}\rangle.\end{aligned}$$

Notice that

$$\mathbb{E}(Z_{t,s} \otimes Z_{t,s})_{i_1,i_2,i_3,i_4} = \begin{cases} 3(e^{s-t}-1)^2, & i_1 = i_2 = i_3 = i_4, \\ (e^{s-t}-1)^2, & i_1 \ne i_2,\ (i_1,i_2) = (i_3,i_4) \text{ or } (i_1,i_2) = (i_4,i_3), \\ 0, \quad \text{else.} \end{cases}$$

So we can bound the inner product by

$$\begin{aligned}\langle \mathbb{E}\, M_t \otimes M_t, \mathbb{E}\, Z_{t,s} \otimes Z_{t,s}\rangle &\lesssim (e^{s-t}-1)^2 \left(\sum_{(i_1,i_2)=(i_3,i_4)} + \sum_{(i_1,i_2)=(i_4,i_3)}\right) \mathbb{E}(M_t \otimes M_t)_{i_1,i_2,i_3,i_4} \\ &\lesssim (e^{s-t}-1)^2 \sum_{(i_1,i_2)=(i_3,i_4)} \mathbb{E}(M_t \otimes M_t)_{i_1,i_2,i_3,i_4} \\ &\lesssim (e^{s-t}-1)^2\, \mathbb{E}\,\|M_t\|_F^2 \\ &\lesssim (e^{s-t}-1)^2\, \mathbb{E}\,\|\nabla^2 \log p_t(x_t)\|_F^4 \\ &\lesssim (e^{s-t}-1)^2 \left(\frac{d}{\sigma_t^2}\right)^4.\end{aligned}$$

where the last inequality comes from Lemma C.7. Next, we bound the second term in (18). By the data processing inequality,

$$\begin{aligned}\mathbb{E}\left(\frac{\mathrm{d}\, P_{x_t + a z_{t,s}, z_{t,s}}(x_t, z_{t,s})}{\mathrm{d}\, P_{x_t, z_{t,s}}(x_t, z_{t,s})}\right)^2 &= \mathbb{E}\left(\frac{\mathrm{d}\, P_{x_t + a z_{t,s} | z_{t,s}}(x_t | z_{t,s})}{\mathrm{d}\, P_{x_t | z_{t,s}}(x_t | z_{t,s})}\right)^2 \\ &\le \mathbb{E}\left(\frac{\mathrm{d}\, P_{x_t + a z_{t,s} | z_{t,s}, x_0}(x_t | z_{t,s}, x_0)}{\mathrm{d}\, P_{x_t | z_{t,s}, x_0}(x_t | z_{t,s}, x_0)}\right)^2 \\ &= \mathbb{E}\left(\frac{\mathrm{d}\, P_{x_t + a z_{t,s} | z_{t,s}, x_0}(x_t | z_{t,s}, x_0)}{\mathrm{d}\, P_{x_t | x_0}(x_t | x_0)}\right)^2.\end{aligned}$$

Notice that $x_t + a z_{t,s} | (z_{t,s}, x_0) \sim \mathcal{N}(\alpha_t^{-1} x_0 + a z_{t,s}, \sigma_t^2 I_d)$ and $x_t | x_0 \sim \mathcal{N}(\alpha_t^{-1} x_0, \sigma_t^2 I_d)$. We can compute the chi-squared divergence explicitly:

$$\mathbb{E}\left(\frac{\mathrm{d}\, P_{x_t + a z_{t,s} | z_{t,s}, x_0}(x_t | z_{t,s}, x_0)}{\mathrm{d}\, P_{x_t | x_0}(x_t | x_0)}\right)^2 = \mathbb{E} \exp\left(\frac{a^2 \|z_{t,s}\|^2}{\sigma_t^2}\right)$$

Finally, the condition $\frac{s-t}{\sigma_t^2} \le \frac{1}{Kd}$ implies $e^{s-t} - 1 \lesssim s - t$ and $\frac{e^{s-t}-1}{\sigma_t^2} \lesssim \frac{1}{Kd}$. Thus for large enough K(actually, $K = 1$ is enough),

$$\mathbb{E} \exp\left(\frac{a^2 \|z_{t,s}\|^2}{\sigma_t^2}\right) = \left(1 - 2\frac{a^2(e^{s-t}-1)}{\sigma_t^2}\right)^{-d/2} \lesssim 1.$$

Combining the bound for the first and the second terms of (18), we conclude that

$$\mathbb{E}\left\|\nabla^2 \log p_t(x_t + a z_{t,s}) z_{t,s}\right\|^2 \lesssim \frac{d^2(s-t)}{\sigma_t^4}. \tag{19}$$

Plugging (19) into (17), we complete the proof. □

C.2. Stability of the Lipschitz Constant

In this subsection, we show that if p_0 satisfies the smoothness condition, p_t is also smooth for sufficiently small t. In particular, under Assumption 4, we can choose $t_0 \asymp \frac{1}{L}$ and an absolute constant C such that for any $0 \le t \le t_0$, the Lipschitz constant of $\nabla \log p_t$ is bounded by CL.

Lemma C.9. *Suppose that Assumption 4 holds. If $\sigma_t^2 \le \frac{\alpha_t}{2L}$, we have $\nabla \log p_t$ is $2L\alpha_t^{-1}$-Lipschitz on $\mathbb{R}^d$.*

Proof. Define a density $q(x) \propto p_0(\alpha_t^{-1} x)$. Then $\nabla \log q$ is $\alpha_t^{-1} L$-Lipschitz. Notice that p_t is the Gaussian perturbation of q. Using Lemma E.3, we write the second-order score function of p_t as

$$\nabla^2 \log p_t(x) = \mathbb{E}_{\tilde{q}_{\sigma_t}(y|x)} \nabla^2 \log q(y) + \mathrm{Var}_{\tilde{q}_{\sigma_t}(y|x)}(\nabla \log q(y)),$$

where $\tilde{q}_{\sigma_t}(y|x)$ is the conditional density given by $\tilde{q}_{\sigma_t}(y|x) \propto q(y) \exp\left(\frac{\|x-y\|^2}{2\sigma_t^2}\right)$. When $\sigma_t^2 \le \frac{\alpha_t}{2L}$, the conditional density satisfies $\log \tilde{q}_{\sigma_t}(y|x) = -\frac{y-x}{\sigma_t^2} + \log q$ is $L\alpha_t^{-1}$-strongly concave, thus it satisfies the Poincaré inequality with a constant $\alpha_t L^{-1}$. From Lemma C.10, we obtain

$$\mathrm{Var}_{\tilde{q}_{\sigma_t}(y|x)}(\nabla \log q(y)) \preceq \alpha_t L^{-1} \mathbb{E}_{\tilde{q}_{\sigma_t}(y|x)}(\nabla^2 \log q(y))(\nabla^2 \log q(y))^\top \preceq L\alpha_t^{-1} I_d.$$

Therefore, we have

$$\mathbb{E}_{\tilde{q}_{\sigma_t}(y|x)} \nabla^2 \log q(y) + \mathrm{Var}_{\tilde{q}_{\sigma_t}(y|x)}(\nabla \log q(y)) \preceq 2L\alpha_t^{-1} I_d.$$

Meanwhile,

$$\mathbb{E}_{\tilde{q}_{\sigma_t}(y|x)} \nabla^2 \log q(y) + \mathrm{Var}_{\tilde{q}_{\sigma_t}(y|x)}(\nabla \log q(y)) \succeq -L\alpha_t^{-1} I_d$$

we complete the proof. □

Lemma C.10. *Let P be a probability distribution on $\mathbb{R}^d$ that satisfies a Poincaré inequality with constant C_P. For any function $f \in C^2(\mathrm{supp}(P))$, we have*

$$\mathrm{Var}_P(\nabla f) \preceq C_P \, \mathbb{E}_P(\nabla^2 f)(\nabla^2 f)^\top.$$

Proof. For any vector $a \in \mathbb{R}^d$, we have

$$\begin{aligned} a^\top \mathrm{Var}_P(\nabla f) a &\le \mathrm{Var}_P(a^\top \nabla f) \\ &\le C_P \, \mathbb{E}_P \left\|\nabla (a^\top \nabla f)\right\|^2 \\ &= C_P \, \mathbb{E}_P \left\|(\nabla^2 f) a\right\|^2 \\ &= C_P a^\top \mathbb{E}_P (\nabla^2 f)(\nabla^2 f)^\top a. \end{aligned}$$

We complete the proof. □

D. Proofs for the Main Theorems

Now we follow the discussion in Section C and combine everything together to complete the proof of our main theorems stated in Section 2.

D.1. Proof of Theorem 2.1

Lemma D.1. *For $t_{k-1} \le t \le t_k$, suppose that $\nabla \log p_t$ is L-Lipschitz for $t_{k-1} \le t \le t_k$. If $L \ge 1, h_k \le 1$, we have*

$$\mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_{t_k}(x_{t_k})\right\|^2 \lesssim dL^2(t_k - t)$$

Proof. The space-discretization error is easily bounded by the Lipschitz condition:

$$\begin{aligned}\mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_t(\alpha_{t,t_k}^{-1} x_{t_k})\right\|^2 &\le dL^2\,\mathbb{E}\left\|x_t - \alpha_{t,t_k}^{-1} x_{t_k}\right\|^2 \\ &= dL^2(e^{t_k - t} - 1) \\ &\lesssim dL^2(t_k - t),\end{aligned} \tag{20}$$

where the last inequality is because of $t_k - t \lesssim 1$. Combining Lemma C.6, Lemma E.2, and (20), we have

$$\begin{aligned}&\mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_{t_k}(x_{t_k})\right\|^2 \\ &\lesssim \mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_t(\alpha_{t,t_k}^{-1} x_{t_k})\right\|^2 + \mathbb{E}\left\|\nabla \log p_t(x_t)\right\|^2 (1 - \alpha_{t,t_k}^{-1})^2 \\ &\lesssim dL^2(t_k - t) + dL(t_k - t)^2 \\ &\lesssim dL^2(t_k - t).\end{aligned}$$

We complete the proof. □

Proof of Theorem 2.1. As shown in Section C, the extra terms arising in the discretization error of Euler-Maruyama scheme can be bounded by Lemma C.5, so we only need to consider the exponential integrator scheme. By Proposition C.3, we can bound the KL divergence between p_0 and $\hat{q}_T$ by

$$\mathrm{KL}(p_0 \| \hat{q}_T) \lesssim \mathrm{KL}(p_T \| \gamma_d) + T\epsilon_0^2 + \sum_{k=1}^{N} \int_{t_{k-1}}^{t_k} \mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_{t_k}(x_{t_k})\right\|^2 \mathrm{d}\, t. \tag{21}$$

The first term in (21) is bounded by Lemma C.4. Then, we apply Lemma D.1 to bound the discretization error:

$$\begin{aligned}&\sum_{k=1}^{N} \int_{t_{k-1}}^{t_k} \mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_{t_k}(x_{t_k})\right\|^2 \mathrm{d}\, t \\ &\lesssim \sum_{k=1}^{N} dL^2 \int_{t_{k-1}}^{t_k} (t_k - t)\, \mathrm{d}\, t \\ &\lesssim dL^2 \sum_{k=1}^{N} h_k^2.\end{aligned}$$

For uniform discretization, the above quantity is $\frac{dT^2L^2}{N}$. We complete the proof. □

D.2. Proof of Theorem 2.2

Lemma D.2. *There is a constant K such that the following holds. In the early stopping setting, suppose that the variance function g satisfies $\frac{h_k}{\sigma_{t_{k-1}}^2} \le \frac{1}{Kd}$ for any integer $1 \le k \le N$. Then we have*

$$\sum_{k=1}^{N} \int_{t_{k-1}}^{t_k} \mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_{t_k}(x_{t_k})\right\|^2 \mathrm{d}\, t \lesssim d^2 \sum_{k=1}^{N} \frac{h_k^2}{\sigma_{t_{k-1}}^4}$$

Proof. By Lemma C.6 and Lemma E.2, we have

$$\begin{aligned}&\mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_{t_k}(x_{t_k})\right\|^2 \\ &\lesssim \mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_t(\alpha_{t,t_k}^{-1} x_{t_k})\right\|^2 + \mathbb{E}\left\|\nabla \log p_t(x_t)\right\|^2 (1 - \alpha_{t,t_k}^{-1})^2 \\ &\lesssim \mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_t(\alpha_{t,t_k}^{-1} x_{t_k})\right\|^2 + \frac{d(1 - \alpha_{t,t_k}^{-1})^2}{\sigma_t^2}.\end{aligned} \tag{22}$$

From Lemma C.8 we have

$$\mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_t(\alpha_{t,t_k}^{-1} x_{t_k})\right\|^2 \lesssim \frac{d^2(t_k - t)}{\sigma_t^4}. \tag{23}$$

Noticing that $\frac{h_k}{\sigma_{t_{k-1}}^2} \lesssim \frac{1}{d}$ implies $\frac{(1-\alpha_{t,t_k}^{-1})^2}{\sigma_t^2} \lesssim \frac{t_k - t}{d}$ and combining this with (22) and (23), we conclude that

$$\mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_{t_k}(x_{t_k})\right\|^2 \lesssim \frac{d^2(t_k - t)}{\sigma_t^4}.$$

Therefore,

$$\begin{aligned}
&\sum_{k=1}^{N} \int_{t_{k-1}}^{t_k} \mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_{t_k}(x_{t_k})\right\|^2 \mathrm{d}\, t \\
&\lesssim \sum_{k=1}^{N} \int_{t_{k-1}}^{t_k} \frac{d^2(t_k - t)}{\sigma_t^4} \mathrm{d}\, t \\
&\lesssim \sum_{k=1}^{N} \int_{t_{k-1}}^{t_k} \frac{d^2(t_k - t)}{\sigma_{t_{k-1}}^4} \mathrm{d}\, t \\
&\lesssim d^2 \sum_{k=1}^{N} \frac{h_k^2}{\sigma_{t_{k-1}}^4}.
\end{aligned}$$

We complete the proof. □

Lemma D.3. *If $K \geq 2$, $c \leq \frac{1}{Kd}$, $t_0 = \delta$, $t_N = T$, and $h_k := t_k - t_{k-1} = c\min\{t_k, 1\}$, then $\frac{h_k}{\sigma_{t_k}^2} \lesssim \frac{1}{Kd}$ for $k = 1, \ldots, N$ and*

$$\Pi := \sum_{k=1}^{N} \frac{h_k^2}{\sigma_{t_{k-1}}^4} \lesssim c\left(\log \frac{1}{\delta} + T\right).$$

Proof. Note that $\sigma_t^2 \asymp \max\{1, t\}$. We consider the sum with $t_k \leq 1$ and $t_k > 1$ separately. For $t_k \leq 1$, we have $\frac{G_k}{\sigma_{k-1}^2} \asymp \frac{ct_k}{t_{k-1}} \leq \frac{2}{Kd}$ (when $K \geq 2$, so $\frac{t_k}{t_{k-1}} \leq 2$). Noting that the number of terms in the sum is $\lesssim \log_{1-c}(\delta)$,

$$\sum_{k:t_k \leq 1} \frac{h_k^2}{\min\{t_{k-1}^2, 1\}} = \sum_{k:t_k \leq 1} \frac{c^2 t_k^2}{t_{k-1}^2} \asymp c^2 \log_{1-c}(\delta) \asymp c^2 \frac{\log(1/\delta)}{c} = c\log(1/\delta). \tag{24}$$

For $t_k > 1$, $\frac{h_k}{\min\{t_{k-1}, 1\}} = c \leq \frac{1}{Kd}$ and

$$\sum_{k:t_k > 1} \frac{h_k^2}{\min\{t_{k-1}^2, 1\}} = \sum_{k:t_k > 1} c^2 \lesssim c^2 \cdot \frac{T}{c} = cT. \tag{25}$$

Combining (24) and (25) gives the result. Note the number of steps is

$$N \lesssim \log_{1-c}(\delta) + \frac{T}{c} = \frac{1}{c}(\log \delta + T).$$

□

Proof of Theorem 2.2. As shown in Section C, the extra terms arising in the discretization error of the Euler-Maruyama scheme can be bounded by Lemma C.5, so we only need to consider the exponential integrator scheme. From Proposition C.3 we obtain

$$\mathrm{KL}(p_{t_0} \| \hat{q}_{T-t_0}) \lesssim \mathrm{KL}(p_T \| \gamma_d) + \sum_{k=1}^{N} \int_{t_{k-1}}^{t_k} \mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_{t_k}(x_{t_k})\right\|^2 \mathrm{d}\, t + T\epsilon_0^2. \tag{26}$$

By bounding the first term in (26) with Lemma C.4 and the second term in (26) with Lemma D.2, we obtain (9). Further more, we can further quantify the term $\Pi = \sum_{k=1}^{N} \frac{G_k^2}{\sigma_{t_{k-1}}^4}$ for exponentially decaying (and then constant) step size with Lemma D.3. □

D.3. Proof of Corollary 2.3 and 2.4

Proof of Corollary 2.3. We use the exponentially decreasing step size in Theorem 2.2. We note that $W_2(P, M_\sharp p_\delta) \lesssim \sqrt{d\sigma_\delta^2} \asymp \sqrt{d\delta}$, so choose $\delta \asymp \frac{\epsilon_{\mathrm{W}}^2}{d}$. Choose $T \asymp \log\left(\frac{d+M_2}{\epsilon_{\mathrm{KL}}^2}\right)$. Also choose $c \asymp \frac{\epsilon_{\mathrm{KL}}^2}{d^2\left(T+\log\left(\frac{1}{\delta}\right)\right)} \gtrsim \frac{\epsilon_{\mathrm{KL}}^2}{d^2 \log\left(\frac{(d+M_2)d}{\epsilon_{\mathrm{KL}}^2 \epsilon_{\mathrm{W}}^2}\right)}$. If $\epsilon_0^2 \lesssim \frac{\epsilon_{\mathrm{KL}}^2}{T^2}$, this ensures that all terms are $\lesssim \epsilon_{\mathrm{KL}}^2$. Choosing appropriate implied constants completes the proof. □

Lemma D.4. *Let μ be the standard Gaussian measure on $N(0, I_d)$. Then*

$$\sup_{\mu(A)\le\epsilon} \int_A \|x\|^2 \mu(dx) \le \epsilon\left(2d + 3\ln\left(\frac{1}{\epsilon}\right) + 3\right) = O\left(\epsilon\left(d + \ln\left(\frac{1}{\epsilon}\right)\right)\right).$$

Proof. By the χ^2 tail bound in (Laurent & Massart, 2000), for $t \ge 0$

$$\mu\left(\|X\|^2 \ge 2d + 3t\right) \le \mathbb{P}\left(\|X\|^2 \ge d + 2\sqrt{dt} + 2t\right) \le e^{-t},$$

so $\|X\|^2$ is stochastically dominated by a random variable with cdf $F(y) = 1 - e^{-\frac{y-2d}{3}}$. Then letting P_Y be the measure corresponding to F,

$$\begin{aligned}\sup_{\mu(A)\le\epsilon} \int_A \|x\|^2 \mu(dx) &\le \sup_{P_Y(A)\le\epsilon} \int_A y P_Y(dy) = \int_{2d+3\ln\left(\frac{1}{\epsilon}\right)}^{\infty} y\, dF(y) \\ &= \epsilon\left(2d + 3\ln\left(\frac{1}{\epsilon}\right)\right) + \int_{2d+3\ln\left(\frac{1}{\epsilon}\right)}^{\infty} e^{-\frac{y-2d}{3}} dy = \epsilon\left(2d + 3\ln\left(\frac{1}{\epsilon}\right)\right) + 3\epsilon\end{aligned}$$

□

Proof of Corollary 2.4. Let $p_\delta^{\mathrm{trunc}}$ be the law of $x_\delta^{\mathrm{trunc}} := x_\delta 1_{\{x_\delta \in B_R(0)\}}$ and define $\hat{q}_{T-\delta}^{\mathrm{trunc}}$ similarly. Note that

$$\begin{aligned} W_2(P, M_\sharp \hat{q}_{T-\delta}^{\mathrm{trunc}}) &\le W_2(P, M_\sharp p_\delta) + W_2(M_\sharp p_\delta, M_\sharp \hat{q}_{T-\delta}^{\mathrm{trunc}}) \\ &\lesssim \sqrt{d\delta} + W_2(p_\delta, \hat{q}_{T-\delta}^{\mathrm{trunc}}). \end{aligned} \tag{27}$$

To bound the second term in (27), we consider a coupling $x_\delta \sim p_\delta$ and $\hat{y}_{T-\delta}^{\mathrm{trunc}} \sim \hat{q}_{T-\delta}^{\mathrm{trunc}}$ such that $x_\delta \ne \hat{y}_{T-\delta}^{\mathrm{trunc}}$ with probability ϵ_{TV}, where

$$\begin{aligned} \epsilon_{\mathrm{TV}} := \mathrm{TV}(p_\delta, \hat{q}_{T-\delta}^{\mathrm{trunc}}) &\le \mathrm{TV}(p_\delta^{\mathrm{trunc}}, \hat{q}_{T-\delta}^{\mathrm{trunc}}) + \mathrm{TV}(p_\delta, p_\delta^{\mathrm{trunc}}) && (28) \\ &\le \mathrm{TV}(p_\delta, \hat{q}_{T-\delta}) + \mathrm{TV}(p_\delta, p_\delta^{\mathrm{trunc}}) && (29) \\ &\le \sqrt{\mathrm{KL}(p_\delta \| \hat{q}_{T-\delta})} + \mathbb{P}\left(\|x_\delta\| \ge R\right) && (30) \\ &= \tilde{O}(\epsilon_0) + \mathbb{P}\left(\|x_\delta\| \ge R\right). && (31) \end{aligned}$$

We used the triangle inequality, data processing inequality, and Pinsker's inequality in (28), (29), and (30), respectively. Express $x_\delta = \alpha_\delta x_0 + \sigma_\delta \xi$, where $x_0 \sim P$, $\xi \sim \mathcal{N}(0, I_d)$. Now

$$\begin{aligned} \mathbb{E}\left\|x_\delta - \hat{y}_{T-\delta}^{\mathrm{trunc}}\right\|^2 &\le \sup_{P(A)\le\epsilon_{\mathrm{TV}}} 2\left(\mathbb{E}\left[\left\|\alpha_\delta x_0 - \hat{y}_{T-\delta}^{\mathrm{trunc}}\right\|^2 1_A\right] + \sigma_\delta^2 \mathbb{E}\left[\|\xi\|^2 1_A\right]\right) \\ &\le 2\left((2M_2 + 2R^2)\epsilon_{\mathrm{TV}} + \sigma_\delta^2 \epsilon_{\mathrm{TV}} \cdot O\left(d + \log\left(\frac{1}{\epsilon_{\mathrm{TV}}}\right)\right)\right), \end{aligned} \tag{32}$$

where the second inequality comes from Lemma D.4. Combining (27), (31), (32) and the choice of parameters in (10), we complete the proof. □

D.4. Proof of Theorem 2.5

Proof of Theorem 2.5. As shown in Section C, the extra terms arising in the discretization error of the Euler-Maruyama scheme can be bounded by Lemma C.5, so we only need to consider the exponential integrator scheme. Using Proposition C.3, we obtain

$$\mathrm{KL}(p_0\|\hat{q}_T) \lesssim \mathrm{KL}(p_T\|\gamma_d) + \sum_{k=1}^{N}\int_{t_{k-1}}^{t_k} \mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_{t_k}(x_{t_k})\right\|^2 \mathrm{d}\, t + T\epsilon_0^2. \tag{33}$$

In the right hand side of (33), the first term is directly bounded by Lemma C.4. Thus we only have to consider the second term, which is the discretization error. Let k_0 be the largest index such that $t_{k_0} \le \frac{1}{L}$. By Lemma D.2 and Lemma D.3,

$$\sum_{k=k_0+1}^{N}\int_{t_{k-1}}^{t_k} \mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_{t_k}(x_{t_k})\right\|^2 \mathrm{d}\, t \lesssim d^2 \sum_{k=k_0+1}^{N} \frac{h_k^2}{\sigma_{t_{k-1}}^4} \lesssim d^2 c(\log L + T).$$

The number of steps for this part is $N - k_0 \lesssim \frac{1}{c}(\log L + T)$. Note $k_0 \lesssim \frac{1}{c}$ so by Lemma D.1 and Lemma C.9,

$$\sum_{k=1}^{k_0}\int_{t_{k-1}}^{t_k} \mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_{t_k}(x_{t_k})\right\|^2 \lesssim dL^2 \sum_{k=1}^{k_0} h_k^2 \lesssim dL^2 \cdot \frac{1}{c}\left(\frac{c}{L}\right)^2 = cd.$$

Thus the total discretization error is bounded by

$$\sum_{k=1}^{N}\int_{t_{k-1}}^{t_k} \mathbb{E}\left\|\nabla \log p_t(x_t) - \nabla \log p_{t_k}(x_{t_k})\right\|^2 \mathrm{d}\, t \lesssim d^2 c(\log L + T)$$

and the total number of steps is $N \lesssim \frac{1}{c}(\log L + T)$. Given the number of steps N, we can choose $c = \frac{\log L + T}{N}$; plugging this in gives the bound. We complete the proof. □

E. Lemmas for Computing Score Functions

In this section, we provide some lemmas for the score function, which will be used in our analysis.

Lemma E.1. *Let P be a probability measure on $\mathbb{R}^d$. Consider the Gaussian perturbation of P that admits a density $p_{\mu,\sigma}(x) \propto \int_{\mathbb{R}^d} \exp\left(-\frac{\|x-\mu y\|^2}{2\sigma^2}\right) \mathrm{d}\, P(y)$. Let $\tilde{P}_{\mu,\sigma}(y|x)$ be the conditional probability measure satisfying $\mathrm{d}\, \tilde{P}_{\mu,\sigma}(y|x) \propto \exp\left(-\frac{\|x-\mu y\|^2}{2\sigma^2}\right) \mathrm{d}\, P(y)$.*

1. *If P admits a density $p \in C^1(\mathbb{R}^d)$, we have*

$$\nabla \log p_{\mu,\sigma}(x) = \frac{1}{\mu}\mathbb{E}_{\tilde{P}_{\mu,\sigma}(y|x)} \nabla_y \log p(y).$$

2. *We have*

$$\nabla \log p_{\mu,\sigma}(x) = \mathbb{E}_{\tilde{P}_{\mu,\sigma}(y|x)}\left(\frac{\mu y - x}{\sigma^2}\right).$$

Proof. The first expression is obtained by

$$\begin{aligned}\nabla \log p_{\mu,\sigma}(x) &= \frac{\int_{\mathbb{R}^d} p(y)\nabla_x\left[\exp\left(-\frac{\|x-\mu y\|^2}{2\sigma^2}\right)\right]\mathrm{d}\, y}{\int_{\mathbb{R}^d} p(y)\exp\left(-\frac{\|x-\mu y\|^2}{2\sigma^2}\right)\mathrm{d}\, y} \\ &= -\frac{\int_{\mathbb{R}^d} p(y)\nabla_y\left[\exp\left(-\frac{\|x-\mu y\|^2}{2\sigma^2}\right)\right]\mathrm{d}\, y}{\mu\int_{\mathbb{R}^d} p(y)\exp\left(-\frac{\|x-\mu y\|^2}{2\sigma^2}\right)\mathrm{d}\, y}\end{aligned}$$

$$= \frac{\int_{\mathbb{R}^d} \nabla_y p(y) \exp\left(-\frac{\|x-\mu y\|^2}{2\sigma^2}\right) \mathrm{d}\, y}{\alpha_{t,s} \int_{\mathbb{R}^d} p(y) \exp\left(-\frac{\|x-\mu y\|^2}{2\sigma^2}\right) \mathrm{d}\, y}$$
$$= \frac{1}{\mu} \mathbb{E}_{\tilde{P}_{\mu,\sigma}(y|x)} \nabla_y \log p(y).$$

For the second expression,

$$\nabla \log p_{\mu,\sigma}(x) = \frac{\int_{\mathbb{R}^d} \nabla_x \left[\exp\left(-\frac{\|x-\mu y\|^2}{2\sigma^2}\right)\right] \mathrm{d}\, P(y)}{\int_{\mathbb{R}^d} \exp\left(-\frac{\|x-\mu y\|^2}{2\sigma^2}\right) \mathrm{d}\, P(y)}$$
$$= \frac{\int_{\mathbb{R}^d} \frac{\mu y - x}{\sigma^2} \exp\left(-\frac{\|x-\mu y\|^2}{2\sigma^2}\right) \mathrm{d}\, P(y)}{\int_{\mathbb{R}^d} \exp\left(-\frac{\|x-\mu y\|^2}{2\sigma^2}\right) \mathrm{d}\, P(y)}$$
$$= \mathbb{E}_{\tilde{P}_{\mu,\sigma}(y|x)}\left(\frac{\mu y - x}{\sigma^2}\right).$$ □

Lemma E.2. *Let $p \in C^1(\mathbb{R}^d)$ be a probability density.*

1. *(Chewi et al., 2022) If $\nabla \log p$ is L-Lipchitz, we have*
$$\mathbb{E}_p \|\nabla \log p(x)\|^2 \le dL.$$
2. *If there exists a probability measure Q and $\sigma > 0$ such that $p(x) \propto \int_{\mathbb{R}^d} \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \mathrm{d}\, Q(y)$. then $\mathbb{E}_p \|\nabla \log p(x)\|^2 \le \frac{d}{\sigma^2}$.*

Proof. 1. Using integration by parts, we have

$$\mathbb{E}_p \|\nabla \log p\|^2 = \int_{\mathbb{R}^d} p(x) \|\nabla \log p(x)\|^2 \mathrm{d}\, x$$
$$= \int_{\mathbb{R}^d} \langle \nabla p(x), \nabla \log p(x) \rangle \mathrm{d}\, x$$
$$= \int_{\mathbb{R}^d} p(x) \Delta \log p(x) \mathrm{d}\, x$$
$$\le dL.$$

2. Using Lemma E.1, we rewrite the score function as

$$\nabla \log p(x) = \mathbb{E}_{\tilde{Q}_\sigma(y|x)}\left(\frac{y-x}{\sigma^2}\right),$$

where $\tilde{Q}_\sigma$ is the conditional density $\mathrm{d}\, \tilde{Q}_\sigma(y|x) \propto \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \mathrm{d}\, Q(y)$. Then the second moment of the score function is bounded by

$$\mathbb{E}_p \|\nabla \log p(x)\|^2 = \mathbb{E}_{p(x)} \left\| \mathbb{E}_{\tilde{Q}_\sigma(y|x)}\left(\frac{y-x}{\sigma^2}\right) \right\|^2 \le \mathbb{E}_{p(x)} \mathbb{E}_{\tilde{Q}_\sigma(y|x)} \left\| \frac{y-x}{\sigma^2} \right\|^2 \le \frac{d}{\sigma^2}.$$

□

Lemma E.3. *Let P be a probability measure on $\mathbb{R}^d$. Consider the density of its Gaussian perturbation $p_\sigma(x) \propto \int_{\mathbb{R}^d} \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \mathrm{d}\, P(y)$. Define a conditional probability measure $\tilde{P}_\sigma(y|x)$ as $\mathrm{d}\, \tilde{P}_\sigma(y|x) \propto \exp\left(\frac{\|x-y\|^2}{2\sigma^2}\right) \mathrm{d}\, P(y)$.*

1. *If P admits a density $p \in C^2(\mathbb{R}^d)$, we have*
$$\nabla^2 \log p_\sigma(x) = \mathbb{E}_{\tilde{P}_\sigma(y|x)} \nabla^2 \log p(y) + \mathrm{Var}_{\tilde{P}_\sigma(y|x)}(\nabla \log p(y)).$$
2. *We have*
$$\nabla^2 \log p_\sigma(x) = \mathrm{Var}_{\tilde{P}_\sigma(y|x)} \left(\frac{y}{\sigma^2}\right) - \frac{I_d}{\sigma^2}.$$

Proof. We rewrite the second-order score function as
$$\nabla^2 \log p_\sigma(x) = \frac{\nabla^2 p_\sigma(x)}{p_\sigma(x)} - \nabla \log p_\sigma(x) (\nabla \log p_\sigma(x))^\top.$$
To prove the first expression, we write
$$\begin{aligned}
\frac{\nabla^2 p_\sigma(x)}{p_\sigma(x)} &= \frac{\int p(y) \nabla_x^2 \exp\left(\frac{-\|x-y\|^2}{2\sigma^2}\right) \mathrm{d}\, y}{\int p(y) \exp\left(\frac{-\|x-y\|^2}{2\sigma^2}\right) \mathrm{d}\, y} \\
&= \frac{\int p(y) \nabla_y^2 \exp\left(\frac{-\|x-y\|^2}{2\sigma^2}\right) \mathrm{d}\, y}{\int p(y) \exp\left(\frac{-\|x-y\|^2}{2\sigma^2}\right) \mathrm{d}\, y} \\
&= \frac{\int \exp\left(\frac{-\|x-y\|^2}{2\sigma^2}\right) \nabla_y^2 p(y) \mathrm{d}\, y}{\int p(y) \exp\left(\frac{-\|x-y\|^2}{2\sigma^2}\right) \mathrm{d}\, y} \\
&= \mathbb{E}_{\tilde{P}_\sigma(y|x)} \frac{\nabla_y^2 p(y)}{p(y)},
\end{aligned}$$
It follows from Lemma E.1 that
$$\nabla \log p_\sigma(x) = \mathbb{E}_{\tilde{P}_\sigma(y|x)} \nabla_y \log p(y).$$
Combining the two terms, we arrive at
$$\begin{aligned}
\nabla^2 \log p_\sigma(x) &= \mathbb{E}_{\tilde{P}_\sigma(y|x)} \frac{\nabla_y^2 p(y)}{p(y)} - \mathbb{E}_{\tilde{P}_\sigma(y|x)} \nabla_y \log p(y) \left(\mathbb{E}_{\tilde{P}_\sigma(y|x)} \nabla_y \log p(y)\right)^\top \\
&= \mathbb{E}_{\tilde{P}_\sigma(y|x)} \nabla_y^2 \log p(y) + \mathrm{Var}_{\tilde{P}_\sigma(y|x)}(\nabla \log p(y)).
\end{aligned}$$
To prove the second expression, we note that
$$\begin{aligned}
\frac{\nabla^2 p_\sigma(x)}{p_\sigma(x)} &= \frac{\int \nabla_x^2 \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \mathrm{d}\, P(y)}{\int \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \mathrm{d}\, P(y)} \\
&= \frac{\int \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \left(\frac{(x-y)(x-y)^\top}{\sigma^4} - \frac{I_d}{\sigma^2}\right) \mathrm{d}\, P(y)}{\int \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \mathrm{d}\, P(y)} \\
&= \mathbb{E}_{\tilde{P}_\sigma(y|x)} \left(\frac{(x-y)(x-y)^\top}{\sigma^4} - \frac{I_d}{\sigma^2}\right).
\end{aligned}$$
It follows from Lemma E.1 that
$$\nabla \log p_\sigma(x) = \mathbb{E}_{\tilde{P}_\sigma(y|x)} \left(\frac{y-x}{\sigma^2}\right).$$
Combining the two terms, we have
$$\begin{aligned}
\nabla^2 \log p_\sigma(x) &= \mathbb{E}_{\tilde{P}_\sigma(y|x)} \left(\frac{(x-y)(x-y)^\top}{\sigma^4} - \frac{I_d}{\sigma^2}\right) - \mathbb{E}_{\tilde{P}_\sigma(y|x)} \frac{y-x}{\sigma^2} \left(\mathbb{E}_{\tilde{P}_\sigma(y|x)} \frac{y-x}{\sigma^2}\right)^\top \\
&= \mathrm{Var}_{\tilde{P}_\sigma(y|x)} \left(\frac{y}{\sigma^2}\right) - \frac{I_d}{\sigma^2}.
\end{aligned}$$
□

F. Technical details for Proposition C.3

Proof of Lemma C.1. By the Fokker-Plank equation, the evolution of p_t and q_t is given by

$$\frac{\partial p_t}{\partial t}(x) = \nabla \cdot \left[-F_1(x,t)p_t(x) + \frac{g(t)^2}{2}\nabla p_t(x) \right] \tag{34}$$

$$\frac{\partial q_t}{\partial t}(x) = \nabla \cdot \left[-F_2(x,t)q_t(x) + \frac{g(t)^2}{2}\nabla q_t(x) \right] \tag{35}$$

Then we have

$$\frac{\partial}{\partial t}\mathrm{KL}(p_t\|q_t) = \int \log\frac{p_t}{q_t}\frac{\partial p_t}{\partial t}\,\mathrm{d}\,x - \int \frac{p_t}{q_t}\frac{\partial q_t}{\partial t}\,\mathrm{d}\,x.$$

For the first term,

$$\begin{aligned}
\int \log\frac{p_t}{q_t}\frac{\partial p_t}{\partial t}\,\mathrm{d}\,x &= \int \nabla\cdot\left[-p_t(x)F_1(x,t) + \frac{g(t)^2}{2}\nabla p_t(x)\right]\log\frac{p_t(x)}{q_t(x)}\,\mathrm{d}\,x \\
&= \int \left\langle \nabla\log\frac{p_t(x)}{q_t(x)}, p_t(x)F_1(x,t) - \frac{g(t)^2}{2}\nabla p_t(x)\right\rangle \mathrm{d}\,x \\
&= \int p_t(x)\left\langle F_1(x,t), \nabla\log\frac{p_t(x)}{q_t(x)}\right\rangle \mathrm{d}\,x - \int \frac{g(t)^2}{2}\left\langle \nabla\log\frac{p_t(x)}{q_t(x)}, \nabla p_t(x)\right\rangle \mathrm{d}\,x
\end{aligned}$$

For the second term,

$$\begin{aligned}
\int \frac{p_t}{q_t}\frac{\partial q_t}{\partial t}\,\mathrm{d}\,x &= \int \frac{p_t}{q_t}\nabla\cdot\left[-F_2(x,t)q_t(x) + \frac{g(t)^2}{2}\nabla q_t(x)\right]\mathrm{d}\,x \\
&= \int \left\langle \nabla\frac{p_t}{q_t}, F_2(x,t)q_t(x) - \frac{g(t)^2}{2}\nabla q_t(x)\right\rangle \mathrm{d}\,x \\
&= \int q_t(x)\left\langle \nabla\frac{p_t}{q_t}, F_2(x,t)\right\rangle \mathrm{d}\,x - \frac{g(t)^2}{2}\left\langle \nabla\frac{p_t}{q_t}, \nabla q_t(x)\right\rangle \mathrm{d}\,x.
\end{aligned}$$

Notice that

$$\begin{aligned}
&\int \left\langle \nabla\frac{p_t}{q_t}, \nabla q_t(x)\right\rangle \mathrm{d}\,x - \int \left\langle \nabla\log\frac{p_t}{q_t}, \nabla p_t(x)\right\rangle \mathrm{d}\,x \\
&= \int \left\langle \frac{q_t\nabla p_t - p_t\nabla q_t}{q_t}, \nabla\log q_t\right\rangle \mathrm{d}\,x - \int p_t\left\langle \nabla\log\frac{p_t}{q_t}, \nabla\log p_t(x)\right\rangle \mathrm{d}\,x \\
&= \int p_t\left\langle \nabla\log\frac{p_t}{q_t}, \nabla\log q_t\right\rangle \mathrm{d}\,x - \int p_t\left\langle \nabla\log\frac{p_t}{q_t}, \nabla\log p_t(x)\right\rangle \mathrm{d}\,x \\
&= -J(p_t\|q_t),
\end{aligned}$$

and

$$\begin{aligned}
&\int p_t(x)\left\langle F_1(x,t), \nabla\log\frac{p_t}{q_t}\right\rangle \mathrm{d}\,x - \int q_t(x)\left\langle \nabla\frac{p_t}{q_t}, F_2(x,t)\right\rangle \mathrm{d}\,x \\
&= \int p_t(x)\left\langle F_1(x,t), \nabla\log\frac{p_t}{q_t}\right\rangle \mathrm{d}\,x - \int p_t(x)\left\langle \nabla\log\frac{p_t}{p_t}, F(x,t)\right\rangle \mathrm{d}\,x \\
&= \int p_t(x)\left\langle \nabla\log\frac{q_t}{p_t}, F_1(x,t) - F_2(x,t)\right\rangle \\
&= \mathbb{E}\left[\left\langle F_1(X_t,t) - F_2(X_t,t), \nabla\log\frac{q_t(X_t)}{p_t(X_t)}\right\rangle\right].
\end{aligned}$$

We complete the proof. □

Proof of Lemma C.2(1). The uniqueness and regularity for the discrete interpolation (14) are obvious since the drift term is linear. Now we check the uniqueness and regularity for (13). In fact, the uniqueness of (13) is guaranteed by the local Lipschitz property of $\nabla \log \tilde{p}_t$ (see, e.g., (Karatzas & Shreve, 1991, Chapter 5, Theorem 2.5)) since $\tilde{p}_t \in C^2(\mathbb{R}^d)$ is supported on $\mathbb{R}^d$. For the regularity, we note that

$$\tilde{p}_{t|t'_k}(x|a) = p_{T-t|T-t'_k}(x|a) = \frac{p_{T-t}(x)p_{T-t'_k|T-t}(a|x)}{p_{T-t'_k}(a)},$$

where $p_{t_1|t_2}$ is the conditional density of x_{t_1} given x_{t_2}. Since $p_{T-t'_k|T-t}(a|x)$ has distribution $\mathcal{N}(\alpha_{T-t,T-t'_k}x, (1-\alpha^2_{T-t,T-t'_k})I_d)$, it is smooth for any $a \in \mathbb{R}^d$, and we have $\tilde{p}_{t|t'_k}(x|a) \in C^2(\mathbb{R}^d)$. □

In order to prove Lemma C.2(2), we need the following.

Lemma F.1. *Let P be a probability measure on $\mathbb{R}^d$. Consider the Gaussian perturbation of P that admits a density $p_\sigma(x) \propto \int_{\mathbb{R}^d} \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \mathrm{d}\, P(y)$. Let $\tilde{P}_\sigma(y|x)$ be the conditional probability measure satisfying $\mathrm{d}\, \tilde{P}_{\mu,\sigma}(y|x) \propto \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \mathrm{d}\, P(y)$. For $x \sim p_\sigma$ we have*

$$\|\nabla \log p_\sigma(x)\|_{\psi_2} \lesssim \sqrt{\frac{d}{\sigma^2}}.$$

Proof. By Lemma E.1, we write the score function of p_σ as

$$\nabla \log p_\sigma(x) = \mathbb{E}_{\tilde{P}_\sigma(y|x)}\left(\frac{y-x}{\sigma^2}\right),$$

where $\tilde{P}_\sigma(y|x)$ be the conditional probability measure satisfying $\mathrm{d}\, \tilde{P}_{\mu,\sigma}(y|x) \propto \exp\left(-\frac{\|x-\mu y\|^2}{2\sigma^2}\right) \mathrm{d}\, P(y)$. For any positive integer p, using the fact that $\frac{y-x}{\sigma}$ is distributed as $\mathcal{N}(0, I_d)$ and the power-mean inequality,

$$\mathbb{E}_{p_\sigma} \|\nabla \log p_\sigma(x)\|^p \le \frac{1}{\sigma^p}\, \mathbb{E}\left\|\frac{y-x}{\sigma}\right\|^p \lesssim \sqrt{\frac{pd}{\sigma^2}}.$$

We complete the proof. □

Proof of Lemma C.2(2). Let $\mathbb{P}_{[t'_k,t]}$ and $\mathbb{Q}_{[t'_k,t]}$ denote be the path measure of $(\tilde{x}_s)_{t'_k \le s \le t}$ and $(\hat{y}_s)_{t'_k \le s \le t}$. For any $a \in \mathbb{R}^d$ we have

$$\mathrm{KL}(\tilde{p}_{t|t'_k}(\cdot|a)\|\hat{q}_{t|t'_k}(\cdot|a)) \le \mathrm{KL}(\mathbb{P}_{[t'_k,t]}(\cdot|\tilde{x}_{t'_k} = a)\|\mathbb{Q}_{[t'_k,t]}(\cdot|\hat{y}_{t'_k} = a)).$$

Thus, it suffices to show

$$\lim_{t \to t'_k +} \mathrm{KL}(\mathbb{P}_{[t'_k,t]}(\cdot|\tilde{x}_{t'_k} = a)\|\mathbb{Q}_{[t'_k,t]}(\cdot|\hat{y}_{t'_k} = a)) = 0 \tag{36}$$

for a.e. $a \in \mathbb{R}^d$. For this, we implement Girsanov change of measure on $\mathbb{P}_{[t'_k,t]}(\cdot|\tilde{x}_{t'_k} = a)$ and $\mathbb{Q}_{[t'_k,t]}(\cdot|\hat{y}_{t'_k} = a)$. If Novikov's condition holds for a.e. $a \in \mathbb{R}^d$, Girsanov's theorem yields

$$\mathrm{KL}(\mathbb{P}_{[t'_k,t]}(\cdot|\tilde{x}_{t'_k} = a)\|\mathbb{Q}_{[t'_k,t]}(\cdot|\hat{y}_{t'_k} = a)) = \mathbb{E}\left[\int_{t'_k}^t \|\nabla \log \tilde{p}(\tilde{x}_s) - s(x, t_{N-k})\|^2 |\tilde{x}_{t_k} = a\right]$$

for the exponential integrator scheme, or

$$\begin{aligned}&\mathrm{KL}(\mathbb{P}_{[t'_k,t]}(\cdot|\tilde{x}_{t'_k} = a)\|\mathbb{Q}_{[t'_k,t]}(\cdot|\hat{y}_{t'_k} = a))\\ &= \mathbb{E}\left[\int_{t'_k}^t \left\|\nabla \log \tilde{p}(\tilde{x}_s) - s(a, t_{N-k}) + \frac{1}{2}(\tilde{x}_s - a)\right\|^2 |\tilde{x}_{t'_k} = a\right]\end{aligned}$$

for the Euler-Maruyama scheme. Hence, (36) is obtained by the Monotone Convergence Theorem and we conclude the proof. Now we check the Novikov condition, which is given by

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\int_{t'_k}^{t}\|\nabla\log\tilde{p}(\tilde{x}_s)-s(a,t_{N-k})\|^2\,\mathrm{d}\,s\right)\Big|\tilde{x}_{t'_k}=a\right]<\infty\qquad\text{(exponential integrator)},$$

or

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\int_{t'_k}^{t}\left\|\nabla\log\tilde{p}(\tilde{x}_s)-s(a,t_{N-k})+\frac{1}{2}(\tilde{x}_s-a)\right\|^2\,\mathrm{d}\,s\right)\Big|\tilde{x}_{t'_k}=a\right]<\infty\,\text{(Euler-Maruyama)}.$$

Hence, it suffices to prove that the following hold for a.e. $a\in\mathbb{R}^d$ when $t-t'_k$ is sufficient small (recall that we only care about the limit $t\to t'_k+$):

$$\mathbb{E}\left[\exp\left(\int_{t'_k}^{t}\|\nabla\log\tilde{p}(\tilde{x}_s)\|^2\,\mathrm{d}\,s\right)|\tilde{x}_{t'_k}=a\right]<\infty \tag{37}$$

$$\mathbb{E}\left[\exp\left(\int_{t'_k}^{t}\|\tilde{x}_s-a\|^2\,\mathrm{d}\,s\right)|\tilde{x}_{t'_k}=a\right]<\infty \tag{38}$$

In fact, by Lemma F.1 we have $\|\nabla\log p_t(x_t)\|_{\psi_2}\lesssim\sqrt{\frac{d}{\sigma_t^2}}$. Thus

$$\left\|\int_{t'_k}^{t}\|\nabla\log\tilde{p}_s(\tilde{x}_s)\|^2\,\mathrm{d}\,s\right\|_{\psi_1}\le\int_{t'_k}^{t}\|\nabla\log\tilde{p}_s(\tilde{x}_s)\|_{\psi_2}^2\,\mathrm{d}\,s\lesssim\frac{d}{\sigma_{T-t}^2}(t-t'_k).$$

When $t-t'_k$ is sufficient small, we have

$$\left\|\int_{t'_k}^{t}\|\nabla\log\tilde{p}_s(\tilde{x}_s)\|^2\right\|_{\psi_1}\le\frac{1}{2},$$

and thus

$$\begin{aligned}&\mathbb{E}_{\tilde{p}_{t'_k}(a)}\left[\mathbb{E}\left[\exp\left(\int_{t'_k}^{t}\|\nabla\log\tilde{p}(\tilde{x}_s)\|^2\,\mathrm{d}\,s\right)|\tilde{x}_{t'_k}=a\right]\right]\\&=\mathbb{E}\left[\exp\left(\int_{t'_k}^{t}\|\nabla\log\tilde{p}(\tilde{x}_s)\|^2\,\mathrm{d}\,s\right)\right]<\infty.\end{aligned}$$

Therefore, (37) holds for a.e. $a\in\mathbb{R}^d$. To verify (38), we split it as

$$\int_{t'_k}^{t}\|\tilde{x}_s-a\|^2\,\mathrm{d}\,s\le 2\int_{t'_k}^{t}\|\tilde{x}_s-\alpha_{T-s,T-t'_k}^{-1}a\|^2+2(\alpha_{T-s,T-t'_k}^{-1}-1)^2(t-t'_k)\|a\|^2. \tag{39}$$

The second term in the right hand side of (39) is a constant so we only need to consider the first term. Note that

$$\begin{aligned}\left\|2\int_{t'_k}^{t}\|\tilde{x}_s-\alpha_{T-s,T-t'_k}^{-1}\tilde{x}_{t'_k}\|^2\,\mathrm{d}\,s\right\|_{\psi_1}&\le 2\int_{t'_k}^{t}\|\tilde{x}_s-\alpha_{T-s,T-t'_k}^{-1}\tilde{x}_{t'_k}\|_{\psi_2}^2\,\mathrm{d}\,s\\&\lesssim 2(e^{-t+t'_k}-1)(t-t'_k).\end{aligned}$$

Thus when $t-t'_k$ is sufficient small we have

$$\left\|2\int_{t'_k}^{t}\|\tilde{x}_s-\alpha_{T-s,T-t'_k}^{-1}\tilde{x}_{t'_k}\|^2\,\mathrm{d}\,s\right\|_{\psi_1}\le\frac{1}{2}$$

and thus

$$\mathbb{E}_{\tilde{p}_{t'_k}(a)}\left[\mathbb{E}\left[\exp\left(2\int_{t'_k}^{t}\|\tilde{x}_s-\alpha_{T-s,T-t'_k}^{-1}a\|^2\,\mathrm{d}\,s\right)|\tilde{x}_{t'_k}=a\right]\right]$$
$$=\mathbb{E}\exp\left(2\int_{t'_k}^{t}\|\tilde{x}_s-\alpha_{T-s,T-t'_k}^{-1}a\|^2\,\mathrm{d}\,s\right)<\infty.$$

We complete the proof of (38). □