Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions

Sitan Chen* Sinho Chewi† Jerry Li‡ Yuanzhi Li§ Adil Salim¶ Anru R. Zhang‖

April 16, 2023

Abstract

We provide theoretical convergence guarantees for score-based generative models (SGMs) such as denoising diffusion probabilistic models (DDPMs), which constitute the backbone of large-scale real-world generative models such as DALL·E 2. Our main result is that, assuming accurate score estimates, such SGMs can efficiently sample from essentially any realistic data distribution. In contrast to prior works, our results (1) hold for an L^2-accurate score estimate (rather than L^∞-accurate); (2) do not require restrictive functional inequality conditions that preclude substantial non-log-concavity; (3) scale polynomially in all relevant problem parameters; and (4) match state-of-the-art complexity guarantees for discretization of the Langevin diffusion, provided that the score error is sufficiently small. We view this as strong theoretical justification for the empirical success of SGMs. We also examine SGMs based on the critically damped Langevin diffusion (CLD). Contrary to conventional wisdom, we provide evidence that the use of the CLD does *not* reduce the complexity of SGMs.

1 Introduction

Score-based generative models (SGMs) are a family of generative models which achieve state-of-the-art performance for generating audio and image data [Soh+15; HJA20; DN21; Kin+21; Son+21a; Son+21b; VKK21]; see, e.g., the recent surveys [Cao+22; Cro+22; Yan+22]. One notable example of an SGM are denoising diffusion probabilistic models (DDPMs) [Soh+15; HJA20], which are a key component in large-scale generative models such as DALL·E 2 [Ram+22]. As the importance of SGMs continues to grow due to newfound applications in commercial domains, it is a pressing question of both practical and theoretical concern to understand the mathematical underpinnings which explain their startling empirical successes.

As we explain in more detail in Section 2, at their mathematical core, SGMs consist of two stochastic processes, which we call the forward process and the reverse process. The forward process transforms samples from a data distribution q (e.g., natural images) into pure noise, whereas the reverse process transforms pure noise into samples from q, hence performing generative modeling. Implementation of the reverse process requires estimation of the *score function* of the law of the forward process, which is typically accomplished by training neural networks on a score matching objective [Hyv05; Vin11; SE19].

Providing precise guarantees for estimation of the score function is difficult, as it requires an understanding of the non-convex training dynamics of neural network optimization that is currently out of reach. However, given the empirical success of neural networks on the score estimation task, a natural and important question is whether or not accurate score estimation implies that SGMs provably converge to the true data distribution in realistic settings. This is a surprisingly delicate question, as even with accurate score estimates, as we

*Department of EECS at University of California, Berkeley, `sitan@seas.harvard.edu`.

†Department of Mathematics at Massachusetts Institute of Technology, `schewi@mit.edu`. Part of this work was done while SC was a research intern at Microsoft Research.

‡Microsoft Research, `jerrl@microsoft.com`.

§Microsoft Research and Machine Learning Department at Carnegie Mellon University, `yuanzhil@andrew.cmu.edu`.

¶Microsoft Research, `adilsalim@microsoft.com`.

‖Departments of Biostatistics & Bioinformatics, Computer Science, Mathematics, and Statistical Science at Duke University, `anru.zhang@duke.edu`.

explain in Section 2.1, there are several other sources of error which could cause the SGM to fail to converge. Indeed, despite a flurry of recent work on this question [BMR22; De +21; De 22; Liu+22; LLT22; Pid22], prior analyses fall short of answering this question, for (at least) one of three main reasons:

1. **Super-polynomial convergence.** The bounds obtained are not quantitative (e.g., [De +21; Liu+22; Pid22]), or scale exponentially in the dimension and other problem parameters [BMR22; De 22], and hence are typically vacuous for the high-dimensional settings of interest in practice.
2. **Strong assumptions on the data distribution.** The bounds require strong assumptions on the true data distribution, such as a log-Sobelev inequality (LSI) (see, e.g., [LLT22]). While the LSI is slightly weaker than log-concavity, it ultimately precludes the presence of substantial non-convexity, which impedes the application of these results to complex and highly multi-modal real-world data distributions. Indeed, obtaining a polynomial-time convergence analysis for SGMs that holds for multi-modal distributions was posed as an open question in [LLT22].
3. **Strong assumptions on the score estimation error.** The bounds require that the score estimate is L^∞-accurate (i.e., *uniformly* accurate), as opposed to L^2-accurate (see, e.g., [De +21]). This is particularly problematic because the score matching objective is an L^2 loss (see Section 2 for details), and there are empirical studies suggesting that in practice, the score estimate is not in fact L^∞-accurate (e.g., [ZC23]). Intuitively, this is because we cannot expect that the score estimate we obtain in practice will be accurate in regions of space where the true density is very low, simply because we do not expect to see many (or indeed, any) samples from such regions.

Providing an analysis which goes beyond these limitations is a pressing first step towards theoretically understanding why SGMs actually work in practice.

Concurrent work. The concurrent and independent work of [LLT23] also obtains similar guarantees to our Corollary 3.

1.1 Our contributions

In this work, we take a step towards bridging theory and practice by providing a convergence guarantee for SGMs, under realistic (in fact, quite minimal) assumptions, which scales polynomially in all relevant problem parameters. Namely, our main result (Theorem 2) only requires the following assumptions on the data distribution q, which we make more quantitative in Section 3:

A1 The score function of the forward process is L-Lipschitz.

A2 The $(2+\eta)$-th moment of q is finite, where $\eta > 0$ is an arbitrarily small constant.

A3 The data distribution q has finite KL divergence w.r.t. the standard Gaussian.

We note that all of these assumptions are either standard or, in the case of **A2**, far weaker than what is needed in prior work. Crucially, unlike prior works, we do *not* assume log-concavity, an LSI, or dissipativity; hence, our assumptions cover *arbitrarily non-log-concave* data distributions. Our main result is summarized informally as follows.

Theorem 1 (informal, see Theorem 2)**.** *Under assumptions* **A1-A3***, and if the score estimation error in* L^2 *is at most* $\widetilde{O}(\varepsilon)$*, then with an appropriate choice of step size, the SGM outputs a measure which is* ε*-close in total variation (TV) distance to* q *in* $\widetilde{O}(L^2 d/\varepsilon^2)$ *iterations.*

We remark that our iteration complexity is actually quite tight: in fact, this matches state-of-the-art discretization guarantees for the Langevin diffusion [VW19; Che+21a].

We find Theorem 1 to be quite surprising, because it shows that SGMs can sample from the data distribution q with polynomial complexity, even when q is highly non-log-concave (a task that is usually intractable), *provided that one has access to an accurate score estimator.* This answers the open question of [LLT22] regarding whether or not SGMs can sample from multimodal distributions, e.g., mixtures of distributions with bounded log-Sobolev constant. In the context of neural networks, our result implies that so long as the neural network succeeds at the learning task, the remaining part of the SGM algorithm based on the diffusion model is principled, in that it admits a strong theoretical justification.

In general, learning the score function is also a difficult task. Nevertheless, our result opens the door to further investigations, such as: do score functions for real-life data have intrinsic (e.g., low-dimensional) structure which can be exploited by neural networks? A positive answer to this question, combined with our sampling result, would then provide an end-to-end guarantee for SGMs.

More generally, our result can be viewed as a black-box reduction of the task of sampling to the task of learning the score function of the forward process, at least for distributions satisfying our mild assumptions. As a simple consequence, existing computational hardness results for learning natural high-dimensional distributions like mixtures of Gaussians [DKS17; Bru+21; GVV22] and pushforwards of Gaussians by shallow ReLU networks [DV21; Che+22a; CLL22] immediately imply hardness of score estimation for these distributions. To our knowledge this yields the first known information-computation gaps for this task.

Arbitrary distributions with bounded support. The assumption that the score function is Lipschitz entails in particular that the data distribution has a density w.r.t. Lebesgue measure; in particular, our theorem fails when q satisfies the manifold hypothesis, i.e., is supported on a lower-dimensional submanifold of $\mathbb{R}^d$. But this is for good reason: it is not possible to obtain non-trivial TV guarantees, because the output distribution of the SGM has full support. Instead, we show in Section 3.2 that we can obtain polynomial convergence guarantees in the bounded Lipschitz metric by stopping the SGM algorithm early, under the *sole* assumption that that data distribution q has bounded support. Since any data distribution encountered in real life satisfies this assumption, our results yield the following compelling takeaway:

Given an L^2-accurate score estimate, SGMs can sample from (essentially) any data distribution.

This constitutes a powerful theoretical justification for the use of SGMs in practice.

Critically damped Langevin diffusion (CLD). Using our techniques, we also investigate the use of the critically damped Langevin diffusion (CLD) for SGMs, which was proposed in [DVK22]. Although numerical experiments and intuition from the log-concave sampling literature suggest that the CLD could potentially speed up sampling via SGMs, we provide theoretical evidence to the contrary: in Section 3.3, we conjecture that SGMs based on the CLD do not exhibit improved dimension dependence compared to the original DDPM algorithm.

1.2 Prior work

We now provide a more detailed comparison to prior work, in addition to the previous discussion above.

By now, there is a vast literature on providing precise complexity estimates for log-concave sampling; see, e.g., the book draft [Che22] for an exposition to recent developments. The proofs in this work build upon the techniques developed in this literature. However, our work addresses the significantly more challenging setting of *non-log-concave* sampling.

The work of [De +21] provides guarantees for the diffusion Schrödinger bridge [Son+21b]. However, as previously mentioned their result is not quantitative, and they require an L^∞-accurate score estimate. The works [BMR22; LLT22][1] instead analyze SGMs under the more realistic assumption of an L^2-accurate score estimate. However, the bounds of [BMR22] suffer from the curse of dimensionality, whereas the bounds of [LLT22] require q to satisfy an LSI.

The recent work of [De 22], motivated by the *manifold hypothesis*, considers a different pointwise assumption on the score estimation error which allows the error to blow up at time 0 and at spatial ∞. We discuss the manifold setting in more detail in Section 3.2. Unfortunately, the bounds of [De 22] also scale exponentially in problem parameters such as the manifold diameter.

After the first version of this work appeared online, we became aware of two concurrent and independent works [LLT23; Liu+22] which share similarities with our work. Namely, [LLT23] obtains similar guarantees to our Corollary 3 below, whereas [Liu+22] uses a similar proof technique as our Theorem 2 (albeit without explicit quantitative bounds).

[1] Unfortunately, the current analysis of [BMR22] contains a gap, as Theorem 27 therein is false: an LSI does not imply contraction in the Wasserstein metric for the Langevin diffusion.

We also mention that the use of reversed SDEs for sampling is also implicit in the interpretation of the proximal sampler algorithm [LST21] given in [Che+22b], and the present work can be viewed as expanding upon the theory of [Che+22b] using a different forward channel (the OU process).

2 Background on SGMs

Throughout this paper, given a probability measure p which admits a density w.r.t. Lebesgue measure, we abuse notation and identify it with its density function. Additionally, we will let q denote the data distribution from which we want to generate new samples. We assume that q is a probability measure on $\mathbb{R}^d$ with full support, and that it admits a smooth density. (See, however, Section 3.2 on applications of our results to the case when q does not admit a density, such as the case when q is supported on a lower-dimensional submanifold of $\mathbb{R}^d$.) In this case, we can write the density of q in the form $q = \exp(-U)$, where $U : \mathbb{R}^d \to \mathbb{R}$ is the *potential*.

In this section, we provide a brief exposition to SGMs, following [Son+21b].

2.1 Background on denoising diffusion probabilistic modeling (DDPM)

Forward process. In denoising diffusion probabilistic modeling (DDPM), we start with a forward process, which is a stochastic differential equation (SDE). For clarity, we consider the simplest possible choice, which is the Ornstein–Uhlenbeck (OU) process

$$\mathrm{d}\bar{X}_t = -\bar{X}_t \,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \qquad \bar{X}_0 \sim q\,, \tag{2.1}$$

where $(B_t)_{t\ge 0}$ is a standard Brownian motion in $\mathbb{R}^d$. The OU process is the unique time-homogeneous Markov process which is also a Gaussian process, with stationary distribution equal to the standard Gaussian distribution γ^d on $\mathbb{R}^d$. In practice, it is also common to introduce a positive smooth function $g : \mathbb{R}_+ \to \mathbb{R}$ and consider the time-rescaled OU process

$$\mathrm{d}\bar{X}_t = -g(t)^2\,\bar{X}_t \,\mathrm{d}t + \sqrt{2}\,g(t)\,\mathrm{d}B_t\,, \qquad X_0 \sim q\,, \tag{2.2}$$

but in this work we stick with the choice $g \equiv 1$.

The forward process has the interpretation of transforming samples from the data distribution q into pure noise. From the well-developed theory of Markov diffusions, it is known that if $q_t \coloneqq \mathrm{law}(X_t)$ denotes the law of the OU process at time t, then $q_t \to \gamma^d$ exponentially fast in various divergences and metrics such as the 2-Wasserstein metric W_2; see [BGL14].

Reverse process. If we reverse the forward process (2.1) in time, then we obtain a process that transforms noise into samples from q, which is the aim of generative modeling. In general, suppose that we have an SDE of the form

$$\mathrm{d}\bar{X}_t = b_t(\bar{X}_t)\,\mathrm{d}t + \sigma_t\,\mathrm{d}B_t\,,$$

where $(\sigma_t)_{t\ge 0}$ is a deterministic matrix-valued process. Then, under mild conditions on the process (e.g., [Föl85; Cat+22]), which are satisfied for all processes under consideration in this work, the reverse process also admits an SDE description. Namely, if we fix the terminal time $T > 0$ and set

$$\bar{X}_t^{\leftarrow} \coloneqq \bar{X}_{T-t}\,, \qquad \text{for } t \in [0,T]\,,$$

then the process $(\bar{X}_t^{\leftarrow})_{t\in[0,T]}$ satisfies the SDE

$$\mathrm{d}\bar{X}_t^{\leftarrow} = b_t^{\leftarrow}(\bar{X}_t^{\leftarrow})\,\mathrm{d}t + \sigma_{T-t}\,\mathrm{d}B_t\,,$$

where the backwards drift satisfies the relation

$$b_t + b_{T-t}^{\leftarrow} = \sigma_t\sigma_t^{\mathsf{T}}\nabla \ln q_t\,, \qquad q_t \coloneqq \mathrm{law}(\bar{X}_t)\,. \tag{2.3}$$

Applying this to the forward process (2.1), we obtain the reverse process

$$\mathrm{d}\bar{X}_t^{\leftarrow} = \{\bar{X}_t^{\leftarrow} + 2\,\nabla \ln q_{T-t}(\bar{X}_t^{\leftarrow})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \qquad \bar{X}_0^{\leftarrow} \sim q_T\,, \tag{2.4}$$

where now $(B_t)_{t\in[0,T]}$ is the reversed Brownian motion.[2] Here, $\nabla \ln q_t$ is called the *score function* for q_t.

[2] For ease of notation, we do not distinguish between the forward and the reverse Brownian motions.

Since q (and hence q_t for $t \geq 0$) is not explicitly known, in order to implement the reverse process the score function must be estimated on the basis of samples.

Score matching. In order to estimate the score function $\nabla \ln q_t$, consider minimizing the $L^2(q_t)$ loss over a function class $\mathcal{F}$,

$$\underset{s_t \in \mathcal{F}}{\text{minimize}} \quad \mathbb{E}_{q_t}[\|s_t - \nabla \ln q_t\|^2] \,, \tag{2.5}$$

where $\mathcal{F}$ could be, e.g., a class of neural networks. The idea of score matching, which goes back to [Hyv05; Vin11], is that after applying integration by parts for the Gaussian measure, the problem (2.5) is *equivalent* to the following problem:

$$\underset{s_t \in \mathcal{F}}{\text{minimize}} \quad \mathbb{E}\Big[\Big\|s_t(\bar{X}_t) + \frac{1}{\sqrt{1-\exp(-2t)}}\, Z_t\Big\|^2\Big] \,, \tag{2.6}$$

where $Z_t \sim \mathsf{normal}(0, I_d)$ is independent of $\bar{X}_0$ and $\bar{X}_t = \exp(-t)\,\bar{X}_0 + \sqrt{1-\exp(-2t)}\, Z_t$, in the sense that (2.5) and (2.6) share the same minimizers. We give a self-contained derivation in Appendix A for the sake of completeness. Unlike (2.5), however, the objective in (2.6) can be replaced with an empirical version and estimated on the basis of samples $\bar{X}_0^{(1)}, \dots, \bar{X}_0^{(n)}$ from q, leading to the finite-sample problem

$$\underset{s_t \in \mathcal{F}}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^n \Big\|s_t(\bar{X}_t^{(i)}) + \frac{1}{\sqrt{1-\exp(-2t)}}\, Z_t^{(i)}\Big\|^2 \,, \tag{2.7}$$

where $(Z_t^{(i)})_{i \in [n]}$ are i.i.d. standard Gaussians independent of the data $(\bar{X}_0^{(i)})_{i \in [n]}$. Moreover, if we parameterize the score function as $s_t = -\frac{1}{\sqrt{1-\exp(-2t)}}\, \widehat{z}_t$, then the empirical problem is equivalent to

$$\underset{\widehat{z}_t \in -\sqrt{1-\exp(-2t)}\,\mathcal{F}}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^n \big\|\widehat{z}_t(\bar{X}_t^{(i)}) - Z_t^{(i)}\big\|^2 \,,$$

which has the illuminating interpretation of predicting the added noise $Z_t^{(i)}$ from the noised data $\bar{X}_t^{(i)}$.

We remark that given the objective function (2.5), it is most natural to assume an $L^2(q_t)$ error bound $\mathbb{E}_{q_t}[\|s_t - \nabla \ln q_t\|^2] \leq \varepsilon_{\text{score}}^2$ for the score estimator. If s_t is taken to be the empirical risk minimizer for an appropriate function class, then guarantees for the $L^2(q_t)$ error can be obtained via standard statistical analysis, as was done in [BMR22].

Discretization and implementation. We now discuss the final steps required to obtain an implementable algorithm. First, in the learning phase, given samples $\bar{X}_0^{(1)}, \dots, \bar{X}_0^{(n)}$ from q (e.g., a database of natural images), we train a neural network on the empirical score matching objective (2.7), see [SE19]. Let $h > 0$ be the step size of the discretization; we assume that we have obtained a score estimate s_{kh} of $\nabla \ln q_{kh}$ for each time $k = 0, 1, \dots, N$, where $T = Nh$.

In order to approximately implement the reverse SDE (2.4), we first replace the score function $\nabla \ln q_{T-t}$ with the estimate s_{T-t}. Then, for $t \in [kh, (k+1)h]$ we freeze the value of this coefficient in the SDE at time kh. It yields the new SDE

$$\mathrm{d}X_t^{\leftarrow} = \{X_t^{\leftarrow} + 2\, s_{T-kh}(X_{kh}^{\leftarrow})\}\, \mathrm{d}t + \sqrt{2}\, \mathrm{d}B_t \,, \qquad t \in [kh, (k+1)h] \,. \tag{2.8}$$

Since this is a linear SDE, it can be integrated in closed form; in particular, conditionally on $X_{kh}^{\leftarrow}$, the next iterate $X_{(k+1)h}^{\leftarrow}$ has an explicit Gaussian distribution.

There is one final detail: although the reverse SDE (2.4) should be started at q_T, we do not have access to q_T directly. Instead, taking advantage of the fact that $q_T \approx \gamma^d$, we instead initialize the algorithm at $X_0^{\leftarrow} \sim \gamma^d$, i.e., from pure noise.

Let $p_t := \mathrm{law}(X_t^{\leftarrow})$ denote the law of the algorithm at time t. The goal of this work is to bound $\mathsf{TV}(p_T, q)$, taking into account three sources of error: (1) the estimation of the score function; (2) the discretization of the SDE with step size $h > 0$; and (3) the initialization of the algorithm at γ^d rather than at q_T.

2.2 Background on the critically damped Langevin diffusion (CLD)

The critically damped Langevin diffusion (CLD) is based on the forward process

$$\begin{aligned} \mathrm{d}\bar{X}_t &= -\bar{V}_t \, \mathrm{d}t \,, \\ \mathrm{d}\bar{V}_t &= -(\bar{X}_t + 2\,\bar{V}_t)\,\mathrm{d}t + 2\,\mathrm{d}B_t \,. \end{aligned} \tag{2.9}$$

Compared to the OU process (2.1), this is now a coupled system of SDEs, where we have introduced a new variable $\bar{V}$ representing the velocity process. The stationary distribution of the process is $\boldsymbol{\gamma}^{2d}$, the standard Gaussian measure on phase space $\mathbb{R}^d \times \mathbb{R}^d$, and we initialize at $\bar{X}_0 \sim q$ and $\bar{V}_0 \sim \gamma^d$.

More generally, the CLD (2.9) is an instance of what is referred to as the *kinetic Langevin* or the *underdamped Langevin* process in the sampling literature. In the context of log-concave sampling, the smoother paths of $\bar{X}$ leads to smaller discretization error, thereby furnishing an algorithm with $\widetilde{O}(\sqrt{d}/\varepsilon)$ gradient complexity (as opposed to sampling based on the overdamped Langevin process, which has complexity $\widetilde{O}(d/\varepsilon^2)$), see [Che+18; SL19; DR20; Ma+21]. In the recent paper [DVK22], Dockhorn, Vahdat, and Kreis proposed to use the CLD as the basis for an SGM and they empirically observed improvements over DDPM.

Applying (2.3), the corresponding reverse process is

$$\begin{aligned} \mathrm{d}\bar{X}_t^{\leftarrow} &= -\bar{V}_t^{\leftarrow} \, \mathrm{d}t \,, \\ \mathrm{d}\bar{V}_t^{\leftarrow} &= \big(\bar{X}_t^{\leftarrow} + 2\,\bar{V}_t^{\leftarrow} + 4\,\nabla_v \ln \boldsymbol{q}_{T-t}(\bar{X}_t^{\leftarrow}, \bar{V}_t^{\leftarrow})\big)\,\mathrm{d}t + 2\,\mathrm{d}B_t \,, \end{aligned} \tag{2.10}$$

where $\boldsymbol{q}_t \coloneqq \mathrm{law}(\bar{X}_t, \bar{V}_t)$ is the law of the forward process at time t. Note that the gradient in the score function is only taken w.r.t. the velocity coordinate. Upon replacing the score function with an estimate $\boldsymbol{s}$, we arrive at the algorithm

$$\begin{aligned} \mathrm{d}X_t^{\leftarrow} &= -V_t^{\leftarrow} \, \mathrm{d}t \,, \\ \mathrm{d}V_t^{\leftarrow} &= \big(X_t^{\leftarrow} + 2\,V_t^{\leftarrow} + 4\,\boldsymbol{s}_{T-kh}(X_{kh}^{\leftarrow}, V_{kh}^{\leftarrow})\big)\,\mathrm{d}t + 2\,\mathrm{d}B_t \,, \end{aligned}$$

for $t \in [kh, (k+1)h]$. We provide further background on the CLD in Section 6.1.

3 Results

We now state our assumptions and our main results.

3.1 Results for DDPM

For DDPM, we make the following mild assumptions on the data distribution q.

Assumption 1 (Lipschitz score). *For all $t \geq 0$, the score $\nabla \ln q_t$ is L-Lipschitz.*

Assumption 2 (second moment bound). *For some $\eta > 0$, $\mathbb{E}_q[\|\cdot\|^{2+\eta}]$ is finite. We also write $\mathfrak{m}_2^2 \coloneqq \mathbb{E}_q[\|\cdot\|^2]$ for the second moment of q.*

For technical reasons, we need to assume that q has a finite moment of order strictly bigger than 2, but our quantitative bounds will only depend on the second moment $\mathfrak{m}_2^2$.

Assumption 1 is standard and has been used in the prior works [BMR22; LLT22]. However, unlike [LLT22], we do not assume Lipschitzness of the score estimate. Moreover, unlike [BMR22; De +21], we do not assume any convexity or dissipativity assumptions on the potential U, and unlike [LLT22] we do not assume that q satisfies a log-Sobolev inequality. Hence, our assumptions cover a wide range of highly non-log-concave data distributions. Our proof technique is fairly robust and even Assumption 1 could be relaxed (as well as other extensions, such as considering the time-changed forward process (2.2)), although we focus on the simplest setting in order to better illustrate the conceptual significance of our results.

We also assume a bound on the score estimation error.

Assumption 3 (score estimation error). *For all $k = 1, \ldots, N$,*

$$\mathbb{E}_{q_{kh}}[\|s_{kh} - \nabla \ln q_{kh}\|^2] \leq \varepsilon_{\mathrm{score}}^2 \,.$$

This is the same assumption as in [LLT22], and as discussed in Section 2.1, it is a natural and realistic assumption in light of the derivation of the score matching objective.

Our main result for DDPM is the following theorem.

Theorem 2 (DDPM)**.** *Suppose that Assumptions 1, 2, and 3 hold. Let p_T be the output of the DDPM algorithm (Section 2.1) at time T, and suppose that the step size $h \coloneqq T/N$ satisfies $h \lesssim 1/L$, where $L \ge 1$. Then, it holds that*

$$\mathsf{TV}(p_T, q) \lesssim \underbrace{\sqrt{\mathsf{KL}(q \,\|\, \gamma^d)}\exp(-T)}_{\text{convergence of forward process}} + \underbrace{(L\sqrt{dh} + L\mathfrak{m}_2 h)\sqrt{T}}_{\text{discretization error}} + \underbrace{\varepsilon_{\text{score}}\sqrt{T}}_{\text{score estimation error}}.$$

Proof. See Section 5. □

To interpret this result, suppose that $\mathsf{KL}(q \,\|\, \gamma^d) \le \operatorname{poly}(d)$ and $\mathfrak{m}_2 \le d$. Choosing $T \asymp \log(\mathsf{KL}(q \,\|\, \gamma^d)/\varepsilon)$ and $h \asymp \frac{\varepsilon^2}{L^2 d}$, and hiding logarithmic factors,

$$\mathsf{TV}(p_T, q) \le \widetilde{O}(\varepsilon + \varepsilon_{\text{score}}), \qquad \text{for } N = \widetilde{\Theta}\Big(\frac{L^2 d}{\varepsilon^2}\Big).$$

In particular, in order to have $\mathsf{TV}(p_T, q) \le \varepsilon$, it suffices to have score error $\varepsilon_{\text{score}} \le \widetilde{O}(\varepsilon)$.

We remark that the iteration complexity of $N = \widetilde{\Theta}(\frac{L^2 d}{\varepsilon^2})$ matches state-of-the-art complexity bounds for the Langevin Monte Carlo (LMC) algorithm for sampling under a log-Sobolev inequality (LSI), see [VW19; Che+21a]. This provides some evidence that our discretization bounds are of the correct order, at least with respect to the dimension and accuracy parameters, and without higher-order smoothness assumptions.

3.2 Consequences for arbitrary data distributions with bounded support

We now elaborate upon the implications of our results under the *sole* assumption that the data distribution q is compactly supported, $\operatorname{supp} q \subseteq \mathsf{B}(0, R)$. In particular, we do not assume that q has a smooth density w.r.t. Lebesgue measure, which allows for studying the case when q is supported on a lower-dimensional submanifold of $\mathbb{R}^d$ as in the *manifold hypothesis*. This setting was investigated recently in [De 22].

For this setting, our results do not apply directly because the score function of q is not well-defined and hence Assumption 1 fails to hold. Also, the bound in Theorem 2 has a term involving $\mathsf{KL}(q \,\|\, \gamma^d)$ which is infinite if q is not absolutely continuous w.r.t. γ^d. As pointed out by [De 22], in general we cannot obtain non-trivial guarantees for $\mathsf{TV}(p_T, q)$, because p_T has full support and therefore $\mathsf{TV}(p_T, q) = 1$ under the manifold hypothesis. Nevertheless, we show that we can apply our results using an early stopping technique.

Namely, consider q_t the law of the OU process at a time $t > 0$, initialized at q. Then, we show in Lemma 16 that, if $t \asymp \varepsilon_{W_2}^2/(\sqrt{d}\,(R \vee \sqrt{d}))$ where $0 < \varepsilon_{W_2} \ll \sqrt{d}$, then q_t satisfies Assumption 1 with $L \lesssim dR^2\,(R \vee \sqrt{d})^2/\varepsilon_{W_2}^4$, $\mathsf{KL}(q_t \,\|\, \gamma^d) \le \operatorname{poly}(R, d, 1/\varepsilon)$, and $W_2(q_t, q) \le \varepsilon_{W_2}$. By substituting q by q_t into the result of Theorem 2, we obtain Corollary 3 below.

Taking q_t as the new target corresponds to stopping the algorithm early: instead of running the algorithm backward for a time T, we run the algorithm backward for a time $T - t$ (note that $T - t$ should be a multiple of the step size h).

Corollary 3 (compactly supported data)**.** *Suppose that q is supported on the ball of radius $R \ge 1$. Let $t \asymp \varepsilon_{W_2}^2/(\sqrt{d}\,(R \vee \sqrt{d}))$. Then, the output p_{T-t} of DDPM is $\varepsilon_{\mathrm{TV}}$-close in TV to the distribution q_t, which is ε_{W_2}-close in W_2 to q, provided that the step size h is chosen appropriately according to Theorem 2 and*

$$N = \widetilde{\Theta}\Big(\frac{d^3 R^4\,(R \vee \sqrt{d})^4}{\varepsilon_{\mathrm{TV}}^2\,\varepsilon_{W_2}^8}\Big) \qquad \text{and} \qquad \varepsilon_{\text{score}} \le \widetilde{O}(\varepsilon_{\mathrm{TV}}).$$

Note that the dependencies in this corollary are polynomial in all of the relevant problem parameters, which vastly improves upon the exponential dependencies of [De 22].

Remark. The statement of Corollary 3 can be simplified by observing that both the TV distance and the W_1 distance are upper bounds for the bounded Lipschitz metric

$$\mathsf{d}_{\mathrm{BL}}(\mu,\nu) \coloneqq \sup\Bigl\{\int f\,\mathrm{d}\mu - \int f\,\mathrm{d}\nu \Bigm| f:\mathbb{R}^d \to [-1,1] \text{ is 1-Lipschitz}\Bigr\}.$$

Hence, Corollary 3 can be phrased as a convergence guarantee in the bounded Lipschitz metric, which is known to metrize weak convergence (see [Dud02, Theorem 11.3.3]).

3.3 Results for CLD

In order to state our results for score-based generative modeling based on the CLD, we must first modify Assumptions 1 and 3 accordingly.

Assumption 4. *For all $t \ge 0$, the score $\nabla_v \ln \boldsymbol{q}_t$ is L-Lipschitz.*

Assumption 5. *For all $k = 1,\dots,N$,*

$$\mathbb{E}_{\boldsymbol{q}_{kh}}[\|\boldsymbol{s}_{kh} - \nabla_v \ln \boldsymbol{q}_{kh}\|^2] \le \varepsilon_{\text{score}}^2\,.$$

If we ignore the dependence on L and assume that the score estimate is sufficiently accurate, then the iteration complexity guarantee of Theorem 2 is $N = \widetilde{\Theta}(d/\varepsilon^2)$. On the other hand, recall from Section 2.2 that based on intuition from the literature on log-concave sampling and from empirical findings in [DVK22], we might expect that SGMs based on the CLD have a smaller iteration complexity than DDPM. We prove the following theorem.

Theorem 4 (CLD)**.** *Suppose that Assumptions 2, 4, and 5 hold. Let $\boldsymbol{p}_T$ be the output of the SGM algorithm based on the CLD (Section 2.2) at time T, and suppose that the step size $h \coloneqq T/N$ satisfies $h \lesssim 1/L$, where $L \ge 1$. Then, there is a universal constant $c > 0$ such that*

$$\mathsf{TV}(\boldsymbol{p}_T, q\otimes\gamma^d) \lesssim \underbrace{\sqrt{\mathsf{KL}(q \,\|\, \gamma^d) + \mathsf{FI}(q \,\|\, \gamma^d)}\exp(-cT)}_{\text{convergence of forward process}} + \underbrace{(L\sqrt{dh} + L\mathfrak{m}_2 h)\sqrt{T}}_{\text{discretization error}} + \underbrace{\varepsilon_{\text{score}}\sqrt{T}}_{\text{score estimation error}}$$

where $\mathsf{FI}(q \,\|\, \gamma^d)$ is the relative Fisher information $\mathsf{FI}(q \,\|\, \gamma^d) \coloneqq \mathbb{E}_q[\|\nabla \ln(q/\gamma^d)\|^2]$.

Proof. See Section 6. ☐

Note that the result of Theorem 4 is in fact no better than our guarantee for DDPM in Theorem 2. Although it is possible that this is an artefact of our analysis, we believe that it is in fact fundamental. As we discuss in Remark 6.2, from the form of the reverse process (2.10), the SGM based on CLD lacks a certain property (that the discretization error should only depend on the size of the increment of the X process, not the increments of both the X and V processes) which is crucial for the improved dimension dependence of the CLD over the Langevin diffusion in log-concave sampling. Hence, in general, we conjecture that under our assumptions, SGMs based on the CLD do not achieve a better dimension dependence than DDPM.

We provide evidence for our conjecture via a lower bound. In our proofs of Theorems 2 and 4, we rely on bounding the KL divergence between certain measures on the path space $\mathcal{C}([0,T];\mathbb{R}^d)$ via Girsanov's theorem. The following result lower bounds this KL divergence, even for the setting in which the score estimate is perfect ($\varepsilon_{\text{score}} = 0$) and the data distribution q is the standard Gaussian.

Theorem 5. *Let $\boldsymbol{p}_T$ be the output of the SGM algorithm based on the CLD (Section 2.2) at time T, where the data distribution q is the standard Gaussian γ^d, and the score estimate is exact ($\varepsilon_{\text{score}} = 0$). Suppose that the step size h satisfies $h \le \frac{1}{10}$. Then, for the path measures $\boldsymbol{P}_T$ and $\boldsymbol{Q}_T^{\leftarrow}$ of the algorithm and the continuous-time process (2.10) respectively (see Section 6 for details), it holds that*

$$\mathsf{KL}(\boldsymbol{Q}_T^{\leftarrow} \,\|\, \boldsymbol{P}_T) \ge dhT\,.$$

Proof. See Section 6.5. □

Theorem 5 shows that in order to make the KL divergence between the path measures small, we must take $h \lesssim 1/d$, which leads to an iteration complexity that scales linearly in the dimension d. Theorem 5 is not a proof that SGMs based on the CLD cannot achieve better than linear dimension dependence, as it is possible that the output $\boldsymbol{p}_T$ of the SGM is close to $q \otimes \gamma^d$ even if the path measures are not close, but it rules out the possibility of obtaining a better dimension dependence via our Girsanov-based proof technique. We believe that it provides compelling evidence for our conjecture, i.e., that under our assumptions, the CLD does not improve the complexity of SGMs over DDPM.

We remark that in this section, we have only considered the error arising from discretization of the SDE. It is possible that the score function for the SGM with the CLD is easier to estimate than the score function for DDPM, providing a *statistical* benefit of using the CLD. Indeed, under the manifold hypothesis, the score $\nabla \ln q_t$ for DDPM blows up at $t = 0$, but the score $\nabla_v \ln \boldsymbol{q}_t$ for CLD is well-defined at $t = 0$, and hence may lead to improvements over DDPM. We do not investigate this question here and leave it as future work.

4 Technical overview

We now give a detailed technical overview for the proof for DDPM (Theorem 2). The proof for CLD (Theorem 4) follows along similar lines.

Recall that we must deal with three sources of error: (1) the estimation of the score function; (2) the discretization of the SDE; and (3) the initialization of the reverse process at γ^d rather than at q_T.

First, we ignore the errors (1) and (2), and focus on the error (3). Hence, we consider the continuous-time reverse SDE (2.4), initialized from either γ^d or from q_T. Let the law of the two processes at time t be denoted $\tilde{p}_t$ and q_{T-t} respectively; how fast do these laws diverge away from each other?

The two main ways to study Markov diffusions is via the 2-Wasserstein distance W_2, or via information divergences such as the KL divergence or the χ^2 divergence. In order for the reverse process to be contractive in the W_2 distance, one typically needs some form of log-concavity assumption for the data distribution q. For example, if $\nabla \ln q(x) = -x/\sigma^2$ (i.e., $q \sim \mathsf{normal}(0, \sigma^2 I_d)$), then for the reverse process (2.4) we have

$$\mathrm{d}\bar{X}_T^{\leftarrow} = \{\bar{X}_T^{\leftarrow} + 2\,\nabla \ln q(\bar{X}_T^{\leftarrow})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t = \big(1 - \frac{2}{\sigma^2}\big)\,\bar{X}_T^{\leftarrow}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,.$$

For $\sigma^2 \gg 1$, the coefficient in front of $\bar{X}_T^{\leftarrow}$ is positive; this shows that for times near T, the reverse process is actually *expansive*, rather than contractive. This poses an obstacle for an analysis in W_2. Although it is possible to perform a W_2 analysis using a weaker condition, such as a dissipativity condition, it typically leads to exponential dependence on the problem parameters (e.g., [De 22]).

On the other hand, the situation is different for an information divergence d. By the data-processing inequality, we always have

$$\mathsf{d}(q_{T-t}, \tilde{p}_t) \le \mathsf{d}(q_T, \tilde{p}_0) = \mathsf{d}(q_T, \gamma^d)\,.$$

This motivates studying the processes via information divergences. We remark that the convergence of reversed SDEs has been studied in the context of log-concave sampling in [Che+22b] for the proximal sampler algorithm [LST21], providing the intuition behind these observations.

Next, we consider the score estimation error (1) and the discretization error (2). In order to perform a discretization analysis in KL or χ^2, there are two salient proof techniques. The first is the interpolation method of [VW19] (originally for KL divergence, but extended to χ^2 divergence in [Che+21a]), which is the method used in [LLT22]. The interpolation method writes down a differential inequality for $\partial_t \mathsf{d}(q_{T-t}, p_t)$, which is used to bound $\mathsf{d}(q_{T-(k+1)h}, p_{(k+1)h})$ in terms of $\mathsf{d}(q_{T-kh}, p_{kh})$ and an additional error term. Unfortunately, the analysis of [LLT22] required taking d to be the χ^2 divergence, for which the interpolation method is quite delicate. In particular, the error term is bounded using a log-Sobolev assumption on q, see [Che+21a] for further discussion. Instead, we pursue the second approach, which is to apply Girsanov's theorem from stochastic calculus and to instead bound the divergence between measures on path space; this turns out to be doable using standard techniques. This is because, as noted in [Che+21a], the Girsanov approach is more flexible as it requires less stringent assumptions.[3]

To elaborate, the main difficulty of using the interpolation method with an L^2-accurate score estimate (Assumption 3) is that the score estimation error is controlled by assumption under the law of the *true* process (2.4), but the interpolation analysis requires a control of the score estimation error under the law of the *algorithm* (2.8). Consequently, the work of [LLT22] required an involved change of measure argument in order to relate the errors under the two processes. In contrast, the Girsanov approach allows us to directly work with the score estimation error under the true process (2.4).

Once the correct framework for the analysis has been chosen, the actual Girsanov discretization argument follows established proof techniques. However, the use of Girsanov's theorem requires a technical condition known as *Novikov's condition* (see (5.1) below), which in fact *fails* to hold under under our minimal assumptions. To circumvent this issue, we use a surprisingly delicate truncation argument in which we apply Girsanov's theorem to a sequence of pairs of auxiliary processes which converge in total variation to the processes we care about. This is the most technical part of our proof and it is deferred to Appendix C.[4]

Notation

Stochastic processes and their laws.

- The data distribution is $q = q_0$.
- The forward process (2.1) is denoted $(\bar{X}_t)_{t\in[0,T]}$, and $\bar{X}_t \sim q_t$.
- The reverse process (2.4) is denoted $(\bar{X}^{\leftarrow}_t)_{t\in[0,T]}$, where $\bar{X}^{\leftarrow}_t := \bar{X}_{T-t} \sim q_{T-t}$.
- The SGM algorithm (2.8) is denoted $(X^{\leftarrow}_t)_{t\in[0,T]}$, and $X^{\leftarrow}_t \sim p_t$. Recall that we initialize at $p_0 = \gamma^d$, the standard Gaussian measure.
- The process $(X^{\leftarrow,q_T}_t)_{t\in[0,T]}$ is the same as $(X^{\leftarrow}_t)_{t\in[0,T]}$, except that we initialize this process at q_T rather than at γ^d. We write $X^{\leftarrow,q_T}_t \sim p^{q_T}_t$.

Conventions for Girsanov's theorem. When we apply Girsanov's theorem, it is convenient to instead think about a single stochastic process, which for ease of notation we denote simply via $(X_t)_{t\in[0,T]}$, and we consider different measures over the path space $\mathcal{C}([0,T];\mathbb{R}^d)$.

The two measures we consider over path space are:

- $Q^{\leftarrow}_T$, under which $(X_t)_{t\in[0,T]}$ has the law of the reverse process (2.4); and
- $P^{q_T}_T$, under which $(X_t)_{t\in[0,T]}$ has the law of the SGM algorithm initialized at q_T (corresponding to the process $(X^{\leftarrow,q_T}_t)_{t\in[0,T]}$ defined above).

Other parameters. We recall that $T > 0$ denotes the total time for which we run the forward process; $h > 0$ is the step size of the discretization; $L \geq 1$ is the Lipschitz constant of the score function; $\mathfrak{m}_2^2 := \mathbb{E}_q[\|\cdot\|^2]$ is the second moment under the data distribution; and $\varepsilon_{\text{score}}$ is the L^2 score estimation error.

Notation for CLD. The notational conventions for the CLD are similar; however, we must also consider a velocity variable V. When discussing quantities which involve both position and velocity (e.g., the joint distribution $\boldsymbol{q}_t$ of $(\bar{X}_t, \bar{V}_t)$), we typically use boldface fonts.

5 Proofs for DDPM

5.1 Preliminaries on Girsanov's theorem

First, we recall the statement of Girsanov's theorem.

[3]After the first draft of this work was made available online, we became aware of the concurrent and independent work of [Liu+22] which also uses an approach based on Girsanov's theorem.

[4]We note that in contrast, [Liu+22] assume that Novikov's condition holds at the outset.

Theorem 6 (Girsanov's theorem, [Le 16, Theorem 5.22])**.** *Let P_T and Q_T be two probability measures on path space $\mathcal{C}([0,T];\mathbb{R}^d)$. Suppose that under P_T, the process $(X_t)_{t\in[0,T]}$ follows*

$$\mathrm{d}X_t = \tilde{b}_t\,\mathrm{d}t + \sigma_t\,\mathrm{d}\tilde{B}_t\,,$$

where $\tilde{B}$ is a P_T-Brownian motion, and under Q_T, the process $(X_t)_{t\in[0,T]}$ follows

$$\mathrm{d}X_t = b_t\,\mathrm{d}t + \sigma_t\,\mathrm{d}B_t\,,$$

where B is a Q_T-Brownian motion. We assume that for each $t>0$, σ_t is a $d\times d$ symmetric positive definite matrix. Then, provided that Novikov's condition holds,

$$\mathbb{E}_{Q_T}\exp\Bigl(\frac{1}{2}\int_0^T \lVert\sigma_t^{-1}\,(\tilde{b}_t - b_t)\rVert^2\,\mathrm{d}t\Bigr) < \infty\,, \tag{5.1}$$

we have that

$$\frac{\mathrm{d}P_T}{\mathrm{d}Q_T} = \exp\Bigl(\int_0^T \sigma_t^{-1}\,(\tilde{b}_t - b_t)\,\mathrm{d}B_t - \frac{1}{2}\int_0^T \lVert\sigma_t^{-1}\,(\tilde{b}_t - b_t)\rVert^2\,\mathrm{d}t\Bigr)\,.$$

We would like to apply Girsanov's theorem with $P_T = P_T^{q_T}$, $Q_T = Q_T^{\leftarrow}$, $\tilde{b}_t = X_t + 2\,s_{T-kh}(X_{kh})$ (for $t\in[kh,(k+1)h]$), $b_t = X_t + 2\,\nabla\ln q_{T-t}(X_t)$, and $\sigma_t = \sqrt{2}\,I_d$. Unfortunately, under Assumptions 1, 2, and 3 alone, Novikov's condition need not hold.[5] In Appendix C we give a workaround for this issue by defining a sequence of suitable truncations of the reverse and SGM processes for which Novikov's condition does hold and showing that these truncations converge in an appropriate sense to the processes we care about. In this section however, we will work under the assumption that $P_T^{q_T}$ and $Q_T^{\leftarrow}$ do satisfy Novikov's condition as it leads to a considerably simpler proof.

Corollary 7. *Assuming that $P_T^{q_T}$ and Q_T satisfy Novikov's condition (5.1), it holds that*

$$\mathsf{KL}(Q_T^{\leftarrow}\parallel P_T^{q_T}) = \mathbb{E}_{Q_T^{\leftarrow}}\ln\frac{\mathrm{d}Q_T^{\leftarrow}}{\mathrm{d}P_T^{q_T}} = \sum_{k=0}^{N-1}\mathbb{E}_{Q_T^{\leftarrow}}\int_{kh}^{(k+1)h}\lVert s_{T-kh}(X_{kh}) - \nabla\ln q_{T-t}(X_t)\rVert^2\,\mathrm{d}t\,.$$

5.2 Girsanov discretization argument

We now apply Corollary 7 and bound the discretization error.

Theorem 8 (discretization error for DDPM)**.** *Suppose that Assumptions 1, 2, and 3 hold. Let $Q_T^{\leftarrow}$ and $P_T^{q_T}$ denote the measures on path space corresponding to the reverse process (2.4) and the SGM algorithm with L^2-accurate score estimate initialized at q_T. Assume that $L\geq 1$ and $h\lesssim 1/L$. Then,*

$$\mathsf{TV}(P_T^{q_T}, Q_T^{\leftarrow})^2 \lesssim (\varepsilon_{\text{score}}^2 + L^2 dh + L^2\mathfrak{m}_2^2 h^2)\,T\,.$$

We first give a proof under the assumption that $P_T^{q_T}$ and $Q_T^{\leftarrow}$ satisfy Novikov's condition (5.1). In Appendix C, we show how to lift this assumption and give a fully rigorous proof of this theorem.

Proof. [Proof assuming that Novikov's condition holds] For $t\in[kh,(k+1)h]$, we can decompose

$$\begin{aligned}
&\mathbb{E}_{Q_T^{\leftarrow}}[\lVert s_{T-kh}(X_{kh}) - \nabla\ln q_{T-t}(X_t)\rVert^2] \\
&\quad\lesssim \mathbb{E}_{Q_T^{\leftarrow}}[\lVert s_{T-kh}(X_{kh}) - \nabla\ln q_{T-kh}(X_{kh})\rVert^2] + \mathbb{E}_{Q_T^{\leftarrow}}[\lVert\nabla\ln q_{T-kh}(X_{kh}) - \nabla\ln q_{T-t}(X_{kh})\rVert^2] \\
&\qquad\quad + \mathbb{E}_{Q_T^{\leftarrow}}[\lVert\nabla\ln q_{T-t}(X_{kh}) - \nabla\ln q_{T-t}(X_t)\rVert^2] \\
&\quad\lesssim \varepsilon_{\text{score}}^2 + \mathbb{E}_{Q_T^{\leftarrow}}\Bigl[\Bigl\lVert\nabla\ln\frac{q_{T-kh}}{q_{T-t}}(X_{kh})\Bigr\rVert^2\Bigr] + L^2\,\mathbb{E}_{Q_T^{\leftarrow}}[\lVert X_{kh} - X_t\rVert^2]\,.
\end{aligned} \tag{5.2}$$

[5] In order to check Novikov's condition, we would want X_0 to have sub-Gaussian tails for instance.

We must bound the change in the score function along the forward process. If $S : \mathbb{R}^d \to \mathbb{R}^d$ is the mapping $S(x) \coloneqq \exp(-(t-kh))\, x$, then $q_{T-kh} = S_\# q_{T-t} * \mathsf{normal}(0, 1-\exp(-2\,(t-kh)))$. We can then use [LLT22, Lemma C.12] (or the more general Lemma 13 that we prove in Section 6.4) with $\alpha = \exp(t-kh) = 1 + O(h)$ and $\sigma^2 = 1 - \exp(-2\,(t-kh)) = O(h)$ to obtain

$$\begin{aligned}\Big\|\nabla \ln \frac{q_{T-kh}}{q_{T-t}}(X_{kh})\Big\|^2 &\lesssim L^2 dh + L^2 h^2 \, \|X_{kh}\|^2 + (1+L^2)\, h^2 \, \|\nabla \ln q_{T-t}(X_{kh})\|^2 \\ &\lesssim L^2 dh + L^2 h^2 \, \|X_{kh}\|^2 + L^2 h^2 \, \|\nabla \ln q_{T-t}(X_{kh})\|^2 \qquad (5.3)\end{aligned}$$

where the last line uses $L \ge 1$.

For the last term,

$$\begin{aligned}\|\nabla \ln q_{T-t}(X_{kh})\|^2 &\lesssim \|\nabla \ln q_{T-t}(X_t)\|^2 + \|\nabla \ln q_{T-t}(X_{kh}) - \nabla \ln q_{T-t}(X_t)\|^2 \\ &\lesssim \|\nabla \ln q_{T-t}(X_t)\|^2 + L^2\, \|X_{kh} - X_t\|^2\,, \qquad (5.4)\end{aligned}$$

where the second term above is absorbed into the third term of the decomposition (5.2). Hence,

$$\begin{aligned}\mathbb{E}_{Q_T^{\leftarrow}}[\|s_{T-kh}(X_{kh}) - \nabla \ln q_{T-t}(X_t)\|^2] &\lesssim \varepsilon_{\rm score}^2 + L^2 dh + L^2 h^2 \, \mathbb{E}_{Q_T^{\leftarrow}}[\|X_{kh}\|^2] \\ &\quad + L^2 h^2 \, \mathbb{E}_{Q_T^{\leftarrow}}[\|\nabla \ln q_{T-t}(X_t)\|^2] + L^2 \, \mathbb{E}_{Q_T^{\leftarrow}}[\|X_{kh} - X_t\|^2]\,.\end{aligned}$$

Using the fact that under $Q_T^{\leftarrow}$, the process $(X_t)_{t\in[0,T]}$ is the time reversal of the forward process $(\bar X_t)_{t\in[0,T]}$, we can apply the moment bounds in Lemma 9 and the movement bound in Lemma 10 to obtain

$$\begin{aligned}\mathbb{E}_{Q_T^{\leftarrow}}[\|s_{T-kh}(X_{kh}) - \nabla \ln q_{T-t}(X_t)\|^2] &\lesssim \varepsilon_{\rm score}^2 + L^2 dh + L^2 h^2\,(d + \mathfrak{m}_2^2) + L^3 dh^2 + L^2\,(\mathfrak{m}_2^2 h^2 + dh) \\ &\lesssim \varepsilon_{\rm score}^2 + L^2 dh + L^2 \mathfrak{m}_2^2 h^2\,.\end{aligned}$$

The result now follows from Corollary 7 and Pinsker's inequality. □

5.3 Proof of Theorem 2

We can now conclude our main result.

Proof. [Proof of Theorem 2] We recall the notation from Section 4. By the data processing inequality,

$$\mathsf{TV}(p_T, q) \le \mathsf{TV}(P_T, P_T^{q_T}) + \mathsf{TV}(P_T^{q_T}, Q_T^{\leftarrow}) \le \mathsf{TV}(q_T, \gamma^d) + \mathsf{TV}(P_T^{q_T}, Q_T^{\leftarrow})\,.$$

Using the convergence of the OU process in KL divergence (see, e.g., [BGL14, Theorem 5.2.1]) and applying Theorem 8 for the second term,

$$\mathsf{TV}(p_T, q) \lesssim \sqrt{\mathsf{KL}(q \,\|\, \gamma^d)}\,\exp(-T) + (\varepsilon_{\rm score} + L\sqrt{dh} + L\mathfrak{m}_2 h)\,\sqrt{T}\,,$$

which proves the result. □

5.4 Auxiliary lemmas

In this section, we prove some auxiliary lemmas which are used in the proof of Theorem 2.

Lemma 9 (moment bounds for DDPM)**.** *Suppose that Assumptions 1 and 2 hold. Let $(\bar X_t)_{t\in[0,T]}$ denote the forward process (2.1).*

1. *(moment bound) For all $t \ge 0$,*

$$\mathbb{E}[\|\bar X_t\|^2] \le d \vee \mathfrak{m}_2^2\,.$$

2. *(score function bound) For all $t \ge 0$,*

$$\mathbb{E}[\|\nabla \ln q_t(\bar X_t)\|^2] \le Ld\,.$$

Proof.

1. Along the OU process, we have $\bar X_t \overset{\mathsf{d}}{=} \exp(-t)\,\bar X_0 + \sqrt{1-\exp(-2t)}\,\xi$, where $\xi \sim \mathsf{normal}(0, I_d)$ is independent of $\bar X_0$. Hence,

$$\mathbb{E}[\|\bar X_t\|^2] = \exp(-2t)\,\mathbb{E}[\|X\|^2] + \{1-\exp(-2t)\}\,d \le d \vee \mathfrak{m}_2^2\,.$$

2. This follows from the L-smoothness of $\ln q_t$ (see, e.g., [VW19, Lemma 9]). We give a short proof for the sake of completeness.

   If $\mathcal{L}_t f := \Delta f - \langle \nabla U_t, \nabla f\rangle$ is the generator associated with $q_t \propto \exp(-U_t)$, then

$$0 = \mathbb{E}_{q_t}\,\mathcal{L} U_t = \mathbb{E}_{q_t}\,\Delta U_t - \mathbb{E}_{q_t}[\|\nabla U_t\|^2] \le Ld - \mathbb{E}_{q_t}[\|\nabla U_t\|^2]\,.$$

□

Lemma 10 (movement bound for DDPM)**.** *Suppose that Assumption 2 holds. Let $(\bar X_t)_{t\in[0,T]}$ denote the forward process (2.1). For $0 \le s < t$ with $\delta := t-s$, if $\delta \le 1$, then*

$$\mathbb{E}[\|\bar X_t - \bar X_s\|^2] \lesssim \delta^2 \mathfrak{m}_2^2 + \delta d\,.$$

Proof. We can write

$$\begin{aligned}\mathbb{E}[\|\bar X_t - \bar X_s\|^2] &= \mathbb{E}\Big[\Big\|-\int_s^t \bar X_r\,\mathrm{d}r + \sqrt{2}\,(B_t - B_s)\Big\|^2\Big] \lesssim \delta \int_s^t \mathbb{E}[\|\bar X_r\|^2]\,\mathrm{d}r + \delta d \lesssim \delta^2\,(d + \mathfrak{m}_2^2) + \delta d \\ &\lesssim \delta^2 \mathfrak{m}_2^2 + \delta d\,,\end{aligned}$$

where we used Lemma 9. □

6 Proofs for CLD

6.1 Background on the CLD process

More generally, for the forward process we can introduce a *friction parameter* $\gamma > 0$ and consider

$$\begin{aligned}\mathrm{d}\bar X_t &= \bar V_t\,\mathrm{d}t\,,\\ \mathrm{d}\bar V_t &= -\bar X_t\,\mathrm{d}t - \gamma\,\bar V_t\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}B_t\,.\end{aligned}$$

If we write $\bar{\boldsymbol\theta}_t := (\bar X_t, \bar V_t)$, then the forward process satisfies the linear SDE

$$\mathrm{d}\bar{\boldsymbol\theta}_t = \boldsymbol{A}_\gamma \bar{\boldsymbol\theta}_t\,\mathrm{d}t + \Sigma_\gamma\,\mathrm{d}B_t\,,\qquad \text{where } \boldsymbol{A}_\gamma := \begin{bmatrix} 0 & 1 \\ -1 & -\gamma\end{bmatrix} \text{ and } \Sigma_\gamma := \begin{bmatrix} 0 \\ \sqrt{2\gamma}\end{bmatrix}.$$

The solution to the SDE is given by

$$\bar{\boldsymbol\theta}_t = \exp(t\boldsymbol{A}_\gamma)\,\bar{\boldsymbol\theta}_0 + \int_0^t \exp\{(t-s)\,\boldsymbol{A}_\gamma\}\,\Sigma_\gamma\,\mathrm{d}B_s\,, \tag{6.1}$$

which means that by the Itô isometry,

$$\mathrm{law}(\bar{\boldsymbol\theta}_t) = \exp(t\boldsymbol{A}_\gamma)_\#\,\mathrm{law}(\bar{\boldsymbol\theta}_0) * \mathsf{normal}\Big(0, \int_0^t \exp\{(t-s)\,\boldsymbol{A}_\gamma\}\,\Sigma_\gamma\Sigma_\gamma^{\mathsf T}\exp\{(t-s)\,\boldsymbol{A}_\gamma^{\mathsf T}\}\,\mathrm{d}s\Big)\,.$$

Since $\det \boldsymbol{A}_\gamma = 1$, $\boldsymbol{A}_\gamma$ is always invertible. Moreover, from $\operatorname{tr}\boldsymbol{A}_\gamma = -\gamma$, one can work out that the spectrum of $\boldsymbol{A}_\gamma$ is

$$\mathrm{spec}(\boldsymbol{A}_\gamma) = \Big\{-\frac{\gamma}{2} \pm \sqrt{\frac{\gamma^2}{4} - 1}\Big\}\,.$$

However, $\boldsymbol{A}_\gamma$ is not diagonalizable. The case of $\gamma = 2$ is special, as it corresponds to the case when the spectrum is $\{-1\}$, and it corresponds to the *critically damped case*. Following [DVK22], which advocated for setting $\gamma = 2$, we will also only consider the critically damped case. This also has the advantage of substantially simplifying the calculations.

6.2 Girsanov discretization argument

In order to apply Girsanov's theorem, we introduce the path measures $\boldsymbol{P}_T^{q_T}$ and $\boldsymbol{Q}_T^{\leftarrow}$, under which

$$\begin{aligned} \mathrm{d}X_t &= -V_t\,\mathrm{d}t\,,\\ \mathrm{d}V_t &= \{X_t + 2\,V_t + 4\,\boldsymbol{s}_{T-kh}(X_{kh}, V_{kh})\}\,\mathrm{d}t + 2\,\mathrm{d}B_t\,, \end{aligned}$$

for $t \in [kh, (k+1)h]$, and

$$\begin{aligned} \mathrm{d}X_t &= -V_t\,\mathrm{d}t\,,\\ \mathrm{d}V_t &= \{X_t + 2\,V_t + 4\,\nabla_v \ln \boldsymbol{q}_{T-t}(X_t, V_t)\}\,\mathrm{d}t + 2\,\mathrm{d}B_t\,, \end{aligned}$$

respectively. We can apply Girsanov's theorem with

$$\tilde{b}_t = \begin{bmatrix} -V_t \\ X_t + 2\,V_t + 4\,\boldsymbol{s}_{T-kh}(X_{kh}, V_{kh}) \end{bmatrix}, \qquad b_t = \begin{bmatrix} -V_t \\ X_t + 2\,V_t + 4\,\nabla_v \ln \boldsymbol{q}_{T-t}(X_t, V_t) \end{bmatrix}, \qquad \sigma_t = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}.$$

Here, σ_t is not invertible, so strictly speaking Theorem 6 does not apply; however, since $\tilde{b}_t - b_t \in \operatorname{range} \sigma_t$, the expression $\sigma_t^{-1}(\tilde{b}_t - b_t)$ is still well-defined, and the following corollary can be justified, provided that Novikov's condition holds.

Corollary 11. *Suppose that Novikov's condition (5.1) holds. Then,*

$$\mathsf{KL}(\boldsymbol{Q}_T^{\leftarrow} \,\|\, \boldsymbol{P}_T^{q_T}) = \mathbb{E}_{\boldsymbol{Q}_T^{\leftarrow}} \ln \frac{\mathrm{d}\boldsymbol{Q}_T^{\leftarrow}}{\mathrm{d}\boldsymbol{P}_T^{q_T}} = 2 \sum_{k=0}^{N-1} \mathbb{E}_{\boldsymbol{Q}_T^{\leftarrow}} \int_{kh}^{(k+1)h} \|\boldsymbol{s}_{T-kh}(X_{kh}, V_{kh}) - \nabla_v \ln \boldsymbol{q}_{T-t}(X_t, V_t)\|^2\,\mathrm{d}t\,.$$

Using this, we now aim to prove the following theorem.

Theorem 12 (discretization error for CLD)**.** *Suppose that Assumptions 2, 4, and 5 hold. Let $\boldsymbol{Q}_T^{\leftarrow}$ and $\boldsymbol{P}_T^{q_T}$ denote the measures on path space corresponding to the reverse process (2.10) and the SGM algorithm with L^2-accurate score estimate initialized at $\boldsymbol{q}_T$. Assume that $L \ge 1$ and $h \lesssim 1/L$. Then,*

$$\mathsf{TV}(\boldsymbol{P}_T^{q_T}, \boldsymbol{Q}_T^{\leftarrow})^2 \lesssim (\varepsilon_{\rm score}^2 + L^2 dh + L^2 \mathfrak{m}_2^2 h^2)\,T\,.$$

Similarly to Section 5.2, we prove the theorem assuming that Novikov's condition (5.1) holds. This assumption can be lifted using similar arguments to the ones in Appendix C, and we omit the details.

Proof. [Proof assuming that Novikov's condition holds] For $t \in [kh, (k+1)h]$, we can decompose

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{Q}_T^{\leftarrow}}[\|\mathbf{s}_{T-kh}(X_{kh}, V_{kh}) - \nabla_v \ln \boldsymbol{q}_{T-t}(X_t, V_t)\|^2] \\ &\quad\lesssim \mathbb{E}_{\boldsymbol{Q}_T^{\leftarrow}}[\|\mathbf{s}_{T-kh}(X_{kh}, V_{kh}) - \nabla_v \ln \boldsymbol{q}_{T-kh}(X_{kh}, V_{kh})\|^2] \\ &\qquad\quad + \mathbb{E}_{\boldsymbol{Q}_T^{\leftarrow}}[\|\nabla_v \ln \boldsymbol{q}_{T-kh}(X_{kh}, V_{kh}) - \nabla_v \ln \boldsymbol{q}_{T-t}(X_{kh}, V_{kh})\|^2] \\ &\qquad\quad + \mathbb{E}_{\boldsymbol{Q}_T^{\leftarrow}}[\|\nabla_v \ln \boldsymbol{q}_{T-t}(X_{kh}, V_{kh}) - \nabla_v \ln \boldsymbol{q}_{T-t}(X_t, V_t)\|^2] \\ &\quad\lesssim \varepsilon_{\rm score}^2 + \mathbb{E}_{\boldsymbol{Q}_T^{\leftarrow}}\Big[\Big\|\nabla_v \ln \frac{\boldsymbol{q}_{T-kh}}{\boldsymbol{q}_{T-t}}(X_{kh}, V_{kh})\Big\|^2\Big] + L^2\,\mathbb{E}_{\boldsymbol{Q}_T^{\leftarrow}}[\|(X_{kh}, V_{kh}) - (X_t, V_t)\|^2]\,. \end{aligned} \tag{6.2}$$

The change in the score function is bounded by Lemma 13, which generalizes [LLT22, Lemma C.12]. From the representation (6.1) of the solution to the CLD, we note that

$$\boldsymbol{q}_{T-kh} = (\boldsymbol{M}_0)_{\#}\boldsymbol{q}_{T-t} * \mathsf{normal}(0, \boldsymbol{M}_1)$$

with

$$\begin{aligned} \boldsymbol{M}_0 &= \exp\big((t-kh)\,\boldsymbol{A}_2\big)\,,\\ \boldsymbol{M}_1 &= \int_0^{t-kh} \exp\{(t-kh-s)\,\boldsymbol{A}_2\}\,\Sigma_2\Sigma_2^{\mathsf{T}} \exp\{(t-kh-s)\,\boldsymbol{A}_2^{\mathsf{T}}\}\,\mathrm{d}s\,. \end{aligned}$$

In particular, since $\|\boldsymbol{A}_2\|_{\mathrm{op}} \lesssim 1$, $\|\boldsymbol{A}_2^{-1}\|_{\mathrm{op}} \lesssim 1$, and $\|\Sigma_2\|_{\mathrm{op}} \lesssim 1$ it follows that $\|\boldsymbol{M}_0\|_{\mathrm{op}} = 1 + O(h)$ and $\|\boldsymbol{M}_1\|_{\mathrm{op}} = O(h)$. Substituting this into Lemma 13, we deduce that if $h \lesssim 1/L$, then

$$\begin{aligned}\Big\|\nabla_v \ln \frac{\boldsymbol{q}_{T-kh}}{\boldsymbol{q}_{T-t}}(X_{kh}, V_{kh})\Big\|^2 &\le \Big\|\nabla \ln \frac{\boldsymbol{q}_{T-kh}}{\boldsymbol{q}_{T-t}}(X_{kh}, V_{kh})\Big\|^2 \\ &\lesssim L^2 dh + L^2 h^2 \left(\|X_{kh}\|^2 + \|V_{kh}\|^2\right) + (1+L^2)\, h^2 \,\|\nabla \ln \boldsymbol{q}_{T-t}(X_{kh}, V_{kh})\|^2 \\ &\lesssim L^2 dh + L^2 h^2 \left(\|X_{kh}\|^2 + \|V_{kh}\|^2\right) + L^2 h^2 \,\|\nabla \ln \boldsymbol{q}_{T-t}(X_{kh}, V_{kh})\|^2 \,,\end{aligned}$$

where in the last step we used $L \ge 1$.

For the last term,

$$\|\nabla \ln \boldsymbol{q}_{T-t}(X_{kh}, V_{kh})\|^2 \lesssim \|\nabla \ln \boldsymbol{q}_{T-t}(X_t, V_t)\|^2 + L^2 \,\|(X_{kh}, V_{kh}) - (X_t, V_t)\|^2 \,,$$

where the second term above is absorbed into the third term of the decomposition (6.2). Hence,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{Q}_T^{\leftarrow}}[\|\mathbf{s}_{T-kh}(X_{kh}, V_{kh}) - \nabla_v \ln \boldsymbol{q}_{T-t}(X_t, V_t)\|^2] \lesssim \varepsilon_{\mathrm{score}}^2 + L^2 dh + L^2 h^2 \,\mathbb{E}_{\boldsymbol{Q}_T^{\leftarrow}}[\|X_{kh}\|^2 + \|V_{kh}\|^2] \\ + L^2 h^2 \,\mathbb{E}_{\boldsymbol{Q}_T^{\leftarrow}}[\|\nabla \ln \boldsymbol{q}_{T-t}(X_t, V_t)\|^2] \\ + L^2 \,\mathbb{E}_{\boldsymbol{Q}_T^{\leftarrow}}[\|(X_{kh}, V_{kh}) - (X_t, V_t)\|^2] \,.\end{aligned}$$

By applying the moment bounds in Lemma 14 together with Lemma 15 on the movement of the CLD,

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{Q}_T^{\leftarrow}}[\|\mathbf{s}_{T-kh}(X_{kh}, V_{kh}) - \nabla_v \ln \boldsymbol{q}_{T-t}(X_t, V_t)\|^2] \\ &\quad\lesssim \varepsilon_{\mathrm{score}}^2 + L^2 dh + L^2 h^2 \,(d + \mathfrak{m}_2^2) + L^3 dh^2 + L^2 \,(dh + \mathfrak{m}_2^2 h^2) \\ &\quad\lesssim \varepsilon_{\mathrm{score}}^2 + L^2 dh + L^2 \mathfrak{m}_2^2 h^2 \,.\end{aligned}$$

Together with Corollary 11 and Pinsker's inequality, it completes the proof. □

Remark. We now pause to discuss why the discretization bound above does not improve upon the result for DDPM (Theorem 8). In the context of log-concave sampling, one instead considers the underdamped Langevin process

$$\begin{aligned}\mathrm{d}X_t &= V_t \,, \\ \mathrm{d}V_t &= -\nabla U(X_t)\,\mathrm{d}t - \gamma\, V_t\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}B_t \,,\end{aligned}$$

which is discretized to yield the algorithm

$$\begin{aligned}\mathrm{d}X_t &= V_t \,, \\ \mathrm{d}V_t &= -\nabla U(X_{kh})\,\mathrm{d}t - \gamma\, V_t\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}B_t \,,\end{aligned}$$

for $t \in [kh, (k+1)h]$. Let $\boldsymbol{P}_T$ denote the path measure for the algorithm, and let $\boldsymbol{Q}_T$ denote the path measure for the continuous-time process. After applying Girsanov's theorem, we obtain

$$\mathsf{KL}(\boldsymbol{Q}_T \,\|\, \boldsymbol{P}_T) \asymp \frac{1}{\gamma} \sum_{k=0}^{N-1} \mathbb{E}_{\boldsymbol{Q}_T} \int_{kh}^{(k+1)h} \|\nabla U(X_t) - \nabla U(X_{kh})\|^2 \,\mathrm{d}t \,.$$

In this expression, note that ∇U depends only on the position coordinate. Since the X process is smoother (as we do not add Brownian motion directly to X), the error $\|\nabla U(X_t) - \nabla U(X_{kh})\|^2$ is of size $O(dh^2)$, which allows us to take step size $h \lesssim 1/\sqrt{d}$. This explains why the use of the underdamped Langevin diffusion leads to improved dimension dependence for log-concave sampling.

In contrast, consider the reverse process, in which

$$\mathsf{KL}(\boldsymbol{Q}_T^{\leftarrow} \,\|\, \boldsymbol{P}_T^{q_T}) = 2 \sum_{k=0}^{N-1} \mathbb{E}_{\boldsymbol{Q}_T^{\leftarrow}} \int_{kh}^{(k+1)h} \|\boldsymbol{s}_{T-kh}(X_{kh}, V_{kh}) - \nabla_v \ln \boldsymbol{q}_{T-t}(X_t, V_t)\|^2 \,\mathrm{d}t \,.$$

Since discretization of the reverse process involves the score function, which depends on both X *and* V, the error now involves controlling $\|V_t - V_{kh}\|^2$, which is of size $O(dh)$ (the process V is not very smooth because it includes a Brownian motion component). Therefore, from the form of the reverse process, we may expect that SGMs based on the CLD do not improve upon the dimension dependence of DDPM.

In Section 6.5, we use this observation in order to prove a rigorous lower bound against discretization of SGMs based on the CLD.

6.3 Proof of Theorem 4

Proof. [Proof of Theorem 4] By the data processing inequality,

$$\mathsf{TV}(\boldsymbol{p}_T, \boldsymbol{q}_0) \le \mathsf{TV}(\boldsymbol{P}_T, \boldsymbol{P}_T^{q_T}) + \mathsf{TV}(\boldsymbol{P}_T^{q_T}, \boldsymbol{Q}_T^{\leftarrow}) \le \mathsf{TV}(\boldsymbol{q}_T, \boldsymbol{\gamma}^{2d}) + \mathsf{TV}(\boldsymbol{P}_T^{q_T}, \boldsymbol{Q}_T^{\leftarrow}).$$

In [Ma+21], following the entropic hypocoercivity approach of [Vil09], Ma et al. consider a Lyapunov functional $\mathcal{L}$ which is equivalent to the sum of the KL divergence and the Fisher information,

$$\mathcal{L}(\boldsymbol{\mu} \,\|\, \boldsymbol{\gamma}^{2d}) \asymp \mathsf{KL}(\boldsymbol{\mu} \,\|\, \boldsymbol{\gamma}^{2d}) + \mathsf{FI}(\boldsymbol{\mu} \,\|\, \boldsymbol{\gamma}^{2d}),$$

which decays exponentially fast in time: there exists a universal constant $c > 0$ such that for all $t \ge 0$,

$$\mathcal{L}(\boldsymbol{q}_t \,\|\, \boldsymbol{\gamma}^{2d}) \le \exp(-ct)\, \mathcal{L}(\boldsymbol{q}_0 \,\|\, \boldsymbol{\gamma}^{2d}).$$

Since $\boldsymbol{q}_0 = q \otimes \gamma^d$ and $\boldsymbol{\gamma}^{2d} = \gamma^d \otimes \gamma^d$, then $\mathcal{L}(\boldsymbol{q}_0 \,\|\, \boldsymbol{\gamma}^{2d}) \lesssim \mathsf{KL}(q \,\|\, \gamma^d) + \mathsf{FI}(q \,\|\, \gamma^d)$. By Pinsker's inequality and Theorem 12, we deduce that

$$\mathsf{TV}(\boldsymbol{p}_T, \boldsymbol{q}_0) \lesssim \sqrt{\mathsf{KL}(q \,\|\, \gamma^d) + \mathsf{FI}(q \,\|\, \gamma^d)}\, \exp(-cT) + (\varepsilon_{\text{score}} + L\sqrt{dh} + L\mathfrak{m}_2 h)\, \sqrt{T},$$

which completes the proof. □

6.4 Auxiliary lemmas

We begin with the perturbation lemma for the score function.

Lemma 13 (score perturbation lemma)**.** *Let $0 < \zeta < 1$. Suppose that $\boldsymbol{M}_0, \boldsymbol{M}_1 \in \mathbb{R}^{2d\times 2d}$ are two matrices, where $\boldsymbol{M}_1$ is symmetric. Also, assume that $\|\boldsymbol{M}_0 - \boldsymbol{I}_{2d}\|_{\text{op}} \le \zeta$, so that $\boldsymbol{M}_0$ is invertible. Let $\boldsymbol{q} = \exp(-\boldsymbol{H})$ be a probability density on $\mathbb{R}^{2d}$ such that $\nabla \boldsymbol{H}$ is L-Lipschitz with $L \le \frac{1}{4\,\|\boldsymbol{M}_1\|_{\text{op}}}$. Then, it holds that*

$$\Big\|\nabla \ln \frac{(\boldsymbol{M}_0)_{\#}\boldsymbol{q} * \mathsf{normal}(0, \boldsymbol{M}_1)}{\boldsymbol{q}}(\boldsymbol{\theta})\Big\| \lesssim L\sqrt{\|\boldsymbol{M}_1\|_{\text{op}}\, d} + L\zeta\,\|\boldsymbol{\theta}\| + (\zeta + L\,\|\boldsymbol{M}_1\|_{\text{op}})\,\|\nabla \boldsymbol{H}(\boldsymbol{\theta})\|.$$

Proof. The proof follows along the lines of [LLT22, Lemma C.12]. First, we show that when $\boldsymbol{M}_0 = \boldsymbol{I}_{2d}$, if $L \le \frac{1}{2\,\|\boldsymbol{M}_1\|_{\text{op}}}$ then

$$\Big\|\nabla \ln \frac{\boldsymbol{q} * \mathsf{normal}(0, \boldsymbol{M}_1)}{\boldsymbol{q}}(\boldsymbol{\theta})\Big\| \lesssim L\sqrt{\|\boldsymbol{M}_1\|_{\text{op}}\, d} + L\,\|\boldsymbol{M}_1\|_{\text{op}}\,\|\nabla \boldsymbol{H}(\boldsymbol{\theta})\|. \tag{6.3}$$

Let $\mathcal{S}$ denote the subspace $\mathcal{S} \coloneqq \operatorname{range} \boldsymbol{M}_1$. Then, since

$$\big(\boldsymbol{q} * \mathsf{normal}(0, \boldsymbol{M}_1)\big)(\boldsymbol{\theta}) = \int_{\boldsymbol{\theta}+\mathcal{S}} \exp\big(-\frac{1}{2}\,\langle \boldsymbol{\theta} - \boldsymbol{\theta}', \boldsymbol{M}_1^{-1}\,(\boldsymbol{\theta} - \boldsymbol{\theta}')\rangle\big)\, \boldsymbol{q}(\mathrm{d}\boldsymbol{\theta}'),$$

where $\boldsymbol{M}_1^{-1}$ is well-defined on $\mathcal{S}$, we have

$$\begin{aligned}\Big\|\nabla \ln \frac{\boldsymbol{q} * \mathsf{normal}(0, \boldsymbol{M}_1)}{\boldsymbol{q}}(\boldsymbol{\theta})\Big\| &= \Big\|\frac{\int_{\boldsymbol{\theta}+\mathcal{S}} \nabla \boldsymbol{H}(\boldsymbol{\theta}')\exp(-\frac{1}{2}\,\langle \boldsymbol{\theta} - \boldsymbol{\theta}', \boldsymbol{M}_1^{-1}\,(\boldsymbol{\theta} - \boldsymbol{\theta}')\rangle)\,\boldsymbol{q}(\mathrm{d}\boldsymbol{\theta}')}{\int_{\boldsymbol{\theta}+\mathcal{S}} \exp(-\frac{1}{2}\,\langle \boldsymbol{\theta} - \boldsymbol{\theta}', \boldsymbol{M}_1^{-1}\,(\boldsymbol{\theta} - \boldsymbol{\theta}')\rangle)\,\boldsymbol{q}(\mathrm{d}\boldsymbol{\theta}')} - \nabla \boldsymbol{H}(\boldsymbol{\theta})\Big\| \\ &= \|\mathbb{E}_{\boldsymbol{q}_{\boldsymbol{\theta}}}\nabla \boldsymbol{H} - \nabla \boldsymbol{H}(\boldsymbol{\theta})\|.\end{aligned}$$

Here, $\boldsymbol{q}_{\boldsymbol{\theta}}$ is the measure on $\boldsymbol{\theta} + \mathcal{S}$ such that

$$\boldsymbol{q}_{\boldsymbol{\theta}}(\mathrm{d}\boldsymbol{\theta}') \propto \exp\bigl(-\frac{1}{2}\,\langle \boldsymbol{\theta} - \boldsymbol{\theta}', \boldsymbol{M}_1^{-1}\,(\boldsymbol{\theta} - \boldsymbol{\theta}')\rangle\bigr)\,\boldsymbol{q}(\mathrm{d}\boldsymbol{\theta}')\,.$$

Note that since $L \le \frac{1}{2\,\|\boldsymbol{M}_1\|_{\mathrm{op}}}$, then if we write $\boldsymbol{q}_{\boldsymbol{\theta}}(\boldsymbol{\theta}') \propto \exp(-\boldsymbol{H}_{\boldsymbol{\theta}}(\boldsymbol{\theta}'))$, we have

$$\nabla^2 \boldsymbol{H}_{\boldsymbol{\theta}} \succeq \bigl(\frac{1}{\|\boldsymbol{M}_1\|_{\mathrm{op}}} - L\bigr)\, I_d \succeq \frac{1}{2\,\|\boldsymbol{M}_1\|_{\mathrm{op}}}\, I_d \qquad \text{on } \boldsymbol{\theta} + \mathcal{S}\,.$$

Let $\boldsymbol{\theta}_\star \in \arg\min \boldsymbol{H}_{\boldsymbol{\theta}}$ denote a mode. We bound

$$\|\mathbb{E}_{\boldsymbol{q}_{\boldsymbol{\theta}}} \nabla \boldsymbol{H} - \nabla \boldsymbol{H}(\boldsymbol{\theta})\| \le L\,\mathbb{E}_{\boldsymbol{\theta}' \sim \boldsymbol{q}_{\boldsymbol{\theta}}}\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \le L\,\mathbb{E}_{\boldsymbol{\theta}' \sim \boldsymbol{q}_{\boldsymbol{\theta}}}\|\boldsymbol{\theta}' - \boldsymbol{\theta}_\star\| + L\,\|\boldsymbol{\theta}_\star - \boldsymbol{\theta}\|\,.$$

For the first term, [DKR22, Proposition 2] yields

$$\mathbb{E}_{\boldsymbol{\theta}' \sim \boldsymbol{q}_{\boldsymbol{\theta}}}\|\boldsymbol{\theta}' - \boldsymbol{\theta}_\star\| \le \sqrt{2\,\|\boldsymbol{M}_1\|_{\mathrm{op}}\, d}\,.$$

For the second term, since the mode satisfies $\nabla \boldsymbol{H}(\boldsymbol{\theta}_\star) + \boldsymbol{M}_1^{-1}\,(\boldsymbol{\theta}_\star - \boldsymbol{\theta}) = 0$, we have

$$\|\boldsymbol{\theta}_\star - \boldsymbol{\theta}\| \le \|\boldsymbol{M}_1\|_{\mathrm{op}}\,\|\nabla \boldsymbol{H}(\boldsymbol{\theta}_\star)\| \le L\,\|\boldsymbol{M}_1\|_{\mathrm{op}}\,\|\boldsymbol{\theta}_\star - \boldsymbol{\theta}\| + \|\boldsymbol{M}_1\|_{\mathrm{op}}\,\|\nabla \boldsymbol{H}(\boldsymbol{\theta})\|$$

which is rearranged to yield

$$\|\boldsymbol{\theta}_\star - \boldsymbol{\theta}\| \le 2\,\|\boldsymbol{M}_1\|_{\mathrm{op}}\,\|\nabla \boldsymbol{H}(\boldsymbol{\theta})\|\,.$$

After combining the bounds, we obtain the claimed estimate (6.3).

Next, we consider the case of general $\boldsymbol{M}_0$. We have

$$\Bigl\|\nabla \ln \frac{(\boldsymbol{M}_0)_\# \boldsymbol{q} * \mathsf{normal}(0, \boldsymbol{M}_1)}{\boldsymbol{q}}(\boldsymbol{\theta})\Bigr\| \le \Bigl\|\nabla \ln \frac{(\boldsymbol{M}_0)_\# \boldsymbol{q} * \mathsf{normal}(0, \boldsymbol{M}_1)}{(\boldsymbol{M}_0)_\# \boldsymbol{q}}(\boldsymbol{\theta})\Bigr\| + \Bigl\|\nabla \ln \frac{(\boldsymbol{M}_0)_\# \boldsymbol{q}}{\boldsymbol{q}}(\boldsymbol{\theta})\Bigr\|.$$

We can apply (6.3) with $(\boldsymbol{M}_0)_\# \boldsymbol{q}$ in place of $\boldsymbol{q}$, noting that $(\boldsymbol{M}_0)_\# \boldsymbol{q} \propto \exp(-\boldsymbol{H}')$ for $\boldsymbol{H}' \coloneqq \boldsymbol{H} \circ \boldsymbol{M}_0$ which is L'-smooth for $L' \coloneqq L\,\|\boldsymbol{M}_0\|_{\mathrm{op}}^2 \lesssim L$, to get

$$\begin{aligned}\Bigl\|\nabla \ln \frac{(\boldsymbol{M}_0)_\# \boldsymbol{q} * \mathsf{normal}(0, \boldsymbol{M}_1)}{(\boldsymbol{M}_0)_\# \boldsymbol{q}}(\boldsymbol{\theta})\Bigr\| &\lesssim L\sqrt{\|\boldsymbol{M}_1\|_{\mathrm{op}}\, d} + L\,\|\boldsymbol{M}_1\|_{\mathrm{op}}\,\|\boldsymbol{M}_0 \nabla \boldsymbol{H}(\boldsymbol{M}_0 \boldsymbol{\theta})\| \\ &\lesssim L\sqrt{\|\boldsymbol{M}_1\|_{\mathrm{op}}\, d} + L\,\|\boldsymbol{M}_1\|_{\mathrm{op}}\,\|\nabla \boldsymbol{H}(\boldsymbol{M}_0 \boldsymbol{\theta})\|\,.\end{aligned}$$

Note that

$$\|\nabla \boldsymbol{H}(\boldsymbol{M}_0 \boldsymbol{\theta})\| \le \|\nabla \boldsymbol{H}(\boldsymbol{\theta})\| + L\,\|(\boldsymbol{M}_0 - \boldsymbol{I}_{2d})\,\boldsymbol{\theta}\| \lesssim \|\nabla \boldsymbol{H}(\boldsymbol{\theta})\| + L\zeta\,\|\boldsymbol{\theta}\|\,.$$

We also have

$$\begin{aligned}\Bigl\|\nabla \ln \frac{(\boldsymbol{M}_0)_\# \boldsymbol{q}}{\boldsymbol{q}}(\boldsymbol{\theta})\Bigr\| &= \|\boldsymbol{M}_0 \nabla \boldsymbol{H}(\boldsymbol{M}_0 \boldsymbol{\theta}) - \nabla \boldsymbol{H}(\boldsymbol{\theta})\| \\ &\le \|\boldsymbol{M}_0 \nabla \boldsymbol{H}(\boldsymbol{M}_0 \boldsymbol{\theta}) - \boldsymbol{M}_0 \nabla \boldsymbol{H}(\boldsymbol{\theta})\| + \|\boldsymbol{M}_0 \nabla \boldsymbol{H}(\boldsymbol{\theta}) - \nabla \boldsymbol{H}(\boldsymbol{\theta})\| \\ &\lesssim L\,\|(\boldsymbol{M}_0 - \boldsymbol{I}_{2d})\,\boldsymbol{\theta}\| + \zeta\,\|\nabla \boldsymbol{H}(\boldsymbol{\theta})\| \lesssim L\zeta\,\|\boldsymbol{\theta}\| + \zeta\,\|\nabla \boldsymbol{H}(\boldsymbol{\theta})\|\,.\end{aligned}$$

Combining the bounds,

$$\begin{aligned}\Bigl\|\nabla \ln \frac{(\boldsymbol{M}_0)_\# \boldsymbol{q} * \mathsf{normal}(0, \boldsymbol{M}_1)}{\boldsymbol{q}}(\boldsymbol{\theta})\Bigr\| &\lesssim L\sqrt{\|\boldsymbol{M}_1\|_{\mathrm{op}}\, d} + L\zeta\,(1 + L\,\|\boldsymbol{M}_1\|_{\mathrm{op}})\,\|\boldsymbol{\theta}\| + (\zeta + L\,\|\boldsymbol{M}_1\|_{\mathrm{op}})\,\|\nabla \boldsymbol{H}(\boldsymbol{\theta})\| \\ &\lesssim L\sqrt{\|\boldsymbol{M}_1\|_{\mathrm{op}}\, d} + L\zeta\,\|\boldsymbol{\theta}\| + (\zeta + L\,\|\boldsymbol{M}_1\|_{\mathrm{op}})\,\|\nabla \boldsymbol{H}(\boldsymbol{\theta})\|\end{aligned}$$

so the lemma follows. □

Next, we prove the moment and movement bounds for the CLD.

Lemma 14 (moment bounds for CLD)**.** *Suppose that Assumptions 2 and 4 hold. Let $(\bar X_t, \bar V_t)_{t\in[0,T]}$ denote the forward process (2.9).*

1. *(moment bound) For all $t \ge 0$,*
$$\mathbb{E}[\|(\bar X_t, \bar V_t)\|^2] \lesssim d + \mathfrak{m}_2^2\,.$$
2. *(score function bound) For all $t \ge 0$,*
$$\mathbb{E}[\|\nabla \ln \boldsymbol{q}_t(\bar X_t, \bar V_t)\|^2] \le Ld\,.$$

Proof.

1. We can write
$$\mathbb{E}[\|(\bar X_t, \bar V_t)\|^2] = W_2^2(\boldsymbol{q}_t, \delta_{\mathbf{0}}) \lesssim W_2^2(\boldsymbol{q}_t, \boldsymbol{\gamma}^{2d}) + W_2^2(\boldsymbol{\gamma}^{2d}, \delta_{\mathbf{0}}) \lesssim d + W_2^2(\boldsymbol{q}_t, \boldsymbol{\gamma}^{2d})\,.$$
Next, the coupling argument of [Che+18] shows that the CLD converges exponentially fast in the Wasserstein metric associated to a twisted norm $|||\cdot|||$ which is equivalent (up to universal constants) to the Euclidean norm $\|\cdot\|$. It implies the following result, see, e.g., [Che+18, Lemma 8]:
$$W_2^2(\boldsymbol{q}_t, \boldsymbol{\gamma}^{2d}) \lesssim W_2^2(\boldsymbol{q}, \boldsymbol{\gamma}^{2d}) \lesssim W_2^2(\boldsymbol{q}, \delta_{\mathbf{0}}) + W_2^2(\delta_{\mathbf{0}}, \boldsymbol{\gamma}^{2d}) \lesssim d + \mathfrak{m}_2^2\,.$$
2. The proof is the same as in Lemma 9.

□

Lemma 15 (movement bound for CLD)**.** *Suppose that Assumptions 2 holds. Let $(\bar X_t, \bar V_t)_{t\in[0,T]}$ denote the forward process (2.9). For $0 < s < t$ with $\delta := t - s$, if $\delta \le 1$,*
$$\mathbb{E}[\|(\bar X_t, \bar V_t) - (\bar X_s, \bar V_s)\|^2] \lesssim \delta^2 \mathfrak{m}_2^2 + \delta d\,.$$

Proof. First,
$$\mathbb{E}[\|\bar X_t - \bar X_s\|^2] = \mathbb{E}\Big[\Big\|\int_s^t \bar V_r \,\mathrm{d}r\Big\|^2\Big] \le \delta \int_s^t \mathbb{E}[\|\bar V_r\|^2]\,\mathrm{d}r \lesssim \delta^2\,(d + \mathfrak{m}_2^2)\,,$$
where we used the moment bound in Lemma 14. Next,
$$\begin{aligned}\mathbb{E}[\|\bar V_t - \bar V_s\|^2] &= \mathbb{E}\Big[\Big\|\int_s^t (-\bar X_r - 2\,\bar V_r)\,\mathrm{d}r + 2\,(B_t - B_s)\Big\|^2\Big] \lesssim \delta \int_s^t \mathbb{E}[\|\bar X_r\|^2 + \|\bar V_r\|^2]\,\mathrm{d}r + \delta d \\ &\lesssim \delta^2\,(d + \mathfrak{m}_2^2) + \delta d\,,\end{aligned}$$
where we used Lemma 14 again. □

6.5 Lower bound against CLD

Proof. [Proof of Theorem 5] Since $\boldsymbol{q}_0 = \gamma^d \otimes \gamma^d = \boldsymbol{\gamma}^{2d}$ is stationary for the forward process (2.9), we have $\boldsymbol{q}_t = \boldsymbol{\gamma}^{2d}$ for all $t \ge 0$. In this proof, since the score estimate is perfect and $\boldsymbol{q}_T = \boldsymbol{\gamma}^{2d}$, we simply denote the path measure for the algorithm as $\boldsymbol{P}_T = \boldsymbol{P}_T^{q_T}$. From Girsanov's theorem in the form of Corollary 11 and from $\boldsymbol{s}_{T-kh}(x, v) = \nabla_v \ln \boldsymbol{q}_{T-kh}(x, v) = -v$, we have
$$\mathsf{KL}(\boldsymbol{Q}_T^{\leftarrow} \,\|\, \boldsymbol{P}_T) = 2 \sum_{k=0}^{N-1} \mathbb{E}_{\boldsymbol{Q}_T^{\leftarrow}} \int_{kh}^{(k+1)h} \|V_{kh} - V_t\|^2\,\mathrm{d}t\,. \tag{6.4}$$

To lower bound this quantity, we use the inequality $\|x+y\|^2 \ge \frac{1}{2}\|x\|^2 - \|y\|^2$ to write, for $t \in [kh,(k+1)h]$

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{Q}_T^{\leftarrow}}[\|V_{kh}-V_t\|^2] = \mathbb{E}[\|\bar V_{T-kh}-\bar V_{T-t}\|^2] &= \mathbb{E}\Big[\Big\|\int_{T-t}^{T-kh}\{-\bar X_s - 2\,\bar V_s\}\,\mathrm{d}s + 2\,(B_{T-kh}-B_{T-t})\Big\|^2\Big] \\
&\ge 2\,\mathbb{E}[\|B_{T-kh}-B_{T-t}\|^2] - \mathbb{E}\Big[\Big\|\int_{T-t}^{T-kh}\{-\bar X_s - 2\,\bar V_s\}\,\mathrm{d}s\Big\|^2\Big] \\
&\ge 2d\,(t-kh) - (t-kh)\int_{T-t}^{T-kh}\mathbb{E}[\|\bar X_s + 2\,\bar V_s\|^2]\,\mathrm{d}s \\
&\ge 2d\,(t-kh) - (t-kh)\int_{T-t}^{T-kh}\mathbb{E}[2\,\|\bar X_s\|^2 + 8\,\|\bar V_s\|^2]\,\mathrm{d}s\,.
\end{aligned}$$

Using the fact that $\bar X_s \sim \gamma^d$ and $\bar V_s \sim \gamma^d$ for all $s \in [0,T]$, we can then bound

$$\mathbb{E}_{\boldsymbol{Q}_T^{\leftarrow}}[\|V_{kh}-V_t\|^2] \ge 2d\,(t-kh) - 10d\,(t-kh)^2 \ge d\,(t-kh)\,,$$

provided that $h \le \frac{1}{10}$. Substituting this into (6.4),

$$\mathsf{KL}(\boldsymbol{Q}_T^{\leftarrow} \,\|\, \boldsymbol{P}_T) \ge 2d\sum_{k=0}^{N-1}\int_{kh}^{(k+1)h}(t-kh)^2\,\mathrm{d}t = dh^2 N = dhT\,.$$

This proves the result. □

This lower bound shows that the Girsanov discretization argument of Theorem 12 is essentially tight (except possibly the dependence on L).

7 Conclusion

In this work, we provided the first convergence guarantees for SGMs which hold under realistic assumptions (namely, L^2-accurate score estimation and arbitrarily non-log-concave data distributions) and which scale polynomially in the problem parameters. Our results take a step towards explaining the remarkable empirical success of SGMs, at least under the assumption that the score function is learned with small L^2 error.

The main limitation of this work is that we did not address the question of when the score function can be learned well. In general, studying the non-convex training dynamics of learning the score function via neural networks is challenging, but we believe that the resolution of this problem, even for simple learning tasks, would shed considerable light on SGMs. Together with the results in this paper, it would yield the first end-to-end guarantees for SGMs.

In another direction, and in light of the interpretation of our result as a reduction of the task of sampling to the task of score function estimation, we ask whether there are situations of interest in which it is easier to algorithmically learn the score function (not necessarily via a neural network) than it is to (directly) sample.

Acknowledgments. We thank Sébastien Bubeck, Yongxin Chen, Tarun Kathuria, Holden Lee, Ruoqi Shen, and Kevin Tian for helpful discussions. S. Chen was supported by NSF Award 2103300. S. Chewi was supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program, as well as the NSF TRIPODS program (award DMS-2022448). A. Zhang was supported in part by NSF CAREER-2203741. Yuanzhi Li was supported by NSF CAREER-2145703.

A Derivation of the score matching objective

In this section, we present a self-contained derivation of the score matching objective (2.6) for the reader's convenience. See also [Hyv05; Vin11; SE19].

Recall that the problem is to solve

$$\operatorname*{minimize}_{s_t \in \mathcal{F}} \quad \mathbb{E}_{q_t}[\|s_t - \nabla \ln q_t\|^2]\,.$$

This objective cannot be evaluated, even if we replace the expectation over q_t with an empirical average over samples from q_t. The trick is to use an integration by parts identity to reformulate the objective. Here, C will denote any constant that does not depend on the optimization variable s_t. Expanding the square,

$$\mathbb{E}_{q_t}[\|s_t - \nabla \ln q_t\|^2] = \mathbb{E}_{q_t}[\|s_t\|^2 - 2\,\langle s_t, \nabla \ln q_t\rangle] + C\,.$$

We can rewrite the second term using integration by parts:

$$\begin{aligned}\int \langle s_t, \nabla \ln q_t\rangle\,\mathrm{d}q_t &= \int \langle s_t, \nabla q_t\rangle = -\int (\operatorname{div} s_t)\,\mathrm{d}q_t \\ &= -\iint (\operatorname{div} s_t)\big(\exp(-t)\,x_0 + \sqrt{1-\exp(-2t)}\,z_t\big)\,\mathrm{d}q(x_0)\,\mathrm{d}\gamma^d(z_t)\,,\end{aligned}$$

where $\gamma^d = \mathsf{normal}(0, I_d)$ and we used the explicit form of the law of the OU process at time t. Recall the Gaussian integration by parts identity: for any vector field $v : \mathbb{R}^d \to \mathbb{R}^d$,

$$\int (\operatorname{div} v)\,\mathrm{d}\gamma^d = \int \langle x, v(x)\rangle\,\mathrm{d}\gamma^d(x)\,.$$

Applying this identity,

$$\int \langle s_t, \nabla \ln q_t\rangle\,\mathrm{d}q_t = -\frac{1}{\sqrt{1-\exp(-2t)}} \int \langle z_t, s_t(x_t)\rangle\,\mathrm{d}q(x_0)\,\mathrm{d}\gamma^d(z_t)$$

where $x_t = \exp(-t)\,x_0 + \sqrt{1-\exp(-2t)}\,z_t$. Substituting this in,

$$\begin{aligned}\mathbb{E}_{q_t}[\|s_t - \nabla \ln q_t\|^2] &= \mathbb{E}\Big[\|s_t(X_t)\|^2 + \frac{2}{\sqrt{1-\exp(-2t)}}\,\langle Z_t, s_t(X_t)\rangle\Big] + C \\ &= \mathbb{E}\Big[\Big\|s(X_t) + \frac{1}{\sqrt{1-\exp(-2t)}}\,Z_t\Big\|^2\Big] + C\,,\end{aligned}$$

where $X_0 \sim q$ and $Z_t \sim \gamma^d$ are independent, and $X_t \coloneqq \exp(-t)\,X_0 + \sqrt{1-\exp(-2t)}\,Z_t$.

B Deferred proofs

Lemma 16. *Suppose that* $\operatorname{supp} q \subseteq \mathsf{B}(0,R)$ *where* $R \ge 1$, *and let* q_t *denote the law of the OU process at time* t, *started at* q. *Let* $\varepsilon > 0$ *be such that* $\varepsilon \ll \sqrt{d}$ *and set* $t \asymp \varepsilon^2/(\sqrt{d}\,(R \vee \sqrt{d}))$. *Then,*

1. $W_2(q_t, q) \le \varepsilon$.
2. q_t *satisfies*

$$\mathsf{KL}(q_t \,\|\, \gamma^d) \lesssim \frac{\sqrt{d}\,(R \vee \sqrt{d})^3}{\varepsilon^2}\,.$$

3. *For every* $t' \ge t$, $q_{t'}$ *satisfies Assumption 1 with*

$$L \lesssim \frac{dR^2\,(R \vee \sqrt{d})^2}{\varepsilon^4}\,.$$

Proof.

1. For the OU process (2.1), we have $\bar X_t \coloneqq \exp(-t)\,\bar X_0 + \sqrt{1-\exp(-2t)}\,Z$, where $Z \sim \mathsf{normal}(0, I_d)$ is independent of $\bar X_0$. Hence, for $t \lesssim 1$,

$$W_2^2(q, q_t) \le \mathbb{E}\big[\big\|\big(1-\exp(-t)\big)\,\bar X_0 + \sqrt{1-\exp(-2t)}\,Z\big\|^2\big]$$
$$= \big(1-\exp(-t)\big)^2\,\mathbb{E}[\|\bar X_0\|^2] + \big(1-\exp(-2t)\big)\,d \lesssim R^2 t^2 + dt\,.$$

   We now take $t \lesssim \min\{\varepsilon/R, \varepsilon^2/d\}$ to ensure that $W_2^2(q, q_t) \le \varepsilon^2$. Since $\varepsilon \ll \sqrt{d}$, it suffices to take $t \asymp \varepsilon^2/(\sqrt{d}\,(R \vee \sqrt{d}))$.

2. For this, we use the short-time regularization result in [OV01, Corollary 2], which implies that

$$\mathsf{KL}(q_t \,\|\, \gamma^d) \le \frac{W_2^2(q, \gamma^d)}{4t} \lesssim \frac{W_2^2(q, \delta_0) + W_2^2(\gamma^d, \delta_0)}{t} \lesssim \frac{\sqrt{d}\,(R \vee \sqrt{d})^3}{\varepsilon^2}\,.$$

3. Using [MS22, Lemma 4], along the OU process,

$$\frac{1}{1-\exp(-2t)}\,I_d - \frac{\exp(-2t)\,R^2}{(1-\exp(-2t))^2}\,I_d \preccurlyeq -\nabla^2 \ln q_t(x) \preccurlyeq \frac{1}{1-\exp(-2t)}\,I_d\,.$$

   With our choice of t, it implies

$$\|\nabla^2 \ln q_{t'}\|_{\mathrm{op}} \lesssim \frac{1}{1-\exp(-2t')} \vee \frac{\exp(-2t')\,R^2}{(1-\exp(-2t'))^2} \lesssim \frac{1}{t} \vee \frac{R^2}{t^2} \lesssim \frac{dR^2\,(R \vee \sqrt{d})^2}{\varepsilon^4}\,.$$

□

C Novikov's condition for DDPM

In Section 5.2, we proved the Girsanov discretization bound for DDPM (Theorem 8) under the assumption that the path measures $P_T^{q_T}$ and $Q_T^{\leftarrow}$ satisfy Novikov's condition (5.1). Unfortunately, Novikov's condition fails under Assumptions 1, 2, and 3 alone. In this section, we remedy this by instead applying the Girsanov discretization argument to a sequence of pairs of auxiliary processes defined in place of $P_T^{q_T}$ and $Q_T^{\leftarrow}$.

C.1 Truncation argument

Cutoff function. We first construct a smooth cutoff function for truncating the drift terms in the processes.

Lemma 17. *For any $R > 0$, there is a smooth function $\phi_R : \mathbb{R}^d \to [0,1]$ satisfying:*

1. *$\phi_R(x) = 1$ for all $\|x\| \le R$,*
2. *$\phi_R(x) = 0$ for all $\|x\| \ge 2R$,*
3. *ϕ_R is $O(1/R)$-Lipschitz.*

Proof. Let $\psi : \mathbb{R} \to [0,1]$ be an $O(1)$-Lipschitz and smooth function for which $\psi(z) = 1$ for all $|z| \le 1$, $\psi(z) = 0$ for all $|z| \ge 2$. This can be constructed, for instance, by taking the piecewise linear function

$$\widetilde\psi(z) \coloneqq \begin{cases} 1\,, & \text{if } |z| \le 1 + c\,,\\ 0\,, & \text{if } |z| \ge 2 - c\,,\\ 1 - \frac{1}{1-2c}\,\big|\,|z| - (1+c)\,\big|\,, & \text{otherwise}\,, \end{cases}$$

for any small constant $c \in (0, \frac{1}{2})$ and convolving $\widetilde{\psi}$ with $z \mapsto c^{-1}\phi(z/c)$ for any mollifier ϕ supported on $(-1, 1)$. Now define $\phi_R(x) := \psi(\|x\|/R)$. Parts 1 and 2 of the lemma are immediate. For Lipschitzness, we can calculate for $x \neq 0$

$$\nabla \phi_R(x) = \frac{1}{R}\,\psi'\big(\frac{\|x\|}{R}\big)\,\frac{x}{\|x\|}\,,$$

which has norm at most $|\psi'(\|x\|/R)|/R = O(1/R)$ as desired, whereas $\nabla \phi_R(0) = 0$. □

Lemma 18. *Suppose that $v : \mathbb{R}^d \to \mathbb{R}^d$ is an L-Lipschitz vector field. Then, $x \mapsto \phi_R(x)\,v(x)$ is L'-Lipschitz for $L' \lesssim L + \frac{1}{R}\,\|v(0)\|$.*

Proof. Let $x, y \in \mathbb{R}^d$. If $\|y\| \le 2R$, then

$$\begin{aligned}\|\phi_R(x)\,v(x) - \phi_R(y)\,v(y)\| &\le \phi_R(x)\,\|v(x) - v(y)\| + \|v(y)\|\,|\phi_R(x) - \phi_R(y)| \\ &\le L\,\|x-y\| + O\big(\frac{1}{R}\big)\,(\|v(0)\| + 2LR)\,\|x-y\| \le O\big(L + \frac{1}{R}\,\|v(0)\|\big)\,\|x-y\|\,.\end{aligned}$$

If $\|y\| > 2R$, then because $\phi_R(y) = 0$,

$$\begin{aligned}\|\phi_R(x)\,v(x) - \phi_R(y)\,v(y)\| = \phi_R(x)\,\|v(x)\| &\le (\|v(0)\| + 2LR)\,|\phi_R(x) - \phi_R(y)| \\ &\le O\big(L + \frac{1}{R}\,\|v(0)\|\big)\,\|x-y\|\,,\end{aligned}$$

as claimed. □

L^∞-accurate score estimate. In order to verify Novikov's condition (5.1), similarly to [LLT22] we replace the score function s, which is only L^2-accurate, with an L^∞-accurate score function s^∞. Define the bad set

$$B_t := \{\|s_t - \nabla \ln q_t\| \ge \varepsilon_{\rm score,\infty}\}\,,$$

where $\varepsilon_{\rm score,\infty} > 0$ is a parameter to be chosen later. We define the L^∞-accurate score estimate to be

$$s_t^\infty := s_t\,\mathbb{1}_{B_t^{\rm c}} + \nabla \ln q_t\,\mathbb{1}_{B_t}\,. \tag{C.1}$$

We note that $\|s_t^\infty - \nabla \ln q_t\| \le \varepsilon_{\rm score,\infty}$.

Truncation argument. Now define the following modified processes: for $t \in [kh, (k+1)h]$,

$$\mathrm{d}\bar X_t^{\leftarrow,\infty} = \phi_R(\bar X_t^{\leftarrow,\infty})\,\{\bar X_t^{\leftarrow,\infty} + 2\,\nabla \ln q_{T-t}(\bar X_t^{\leftarrow,\infty})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \tag{C.2}$$

$$\mathrm{d}X_t^{\leftarrow,\infty,q_T} = \phi_R(X_t^{\leftarrow,\infty,q_T})\,\{X_t^{\leftarrow,\infty,q_T} + 2\,s^\infty_{T-kh}(X_{kh}^{\leftarrow,\infty,q_T})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \tag{C.3}$$

where $\bar X_0^{\leftarrow,\infty} = X_0^{\leftarrow,\infty,q_T}$ is obtained by sampling $\bar X_T \sim q_T$ and setting $\bar X_0^{\leftarrow,\infty} = \bar X_T$ if $\|\bar X_T\| \le R$ and setting $\bar X_0^{\leftarrow,\infty} = 0$ otherwise. For clarity of presentation, we suppress the dependence on R, $\varepsilon_{\rm score,\infty}$ in the notation.

Let $q_t^{\leftarrow,\infty} := \mathrm{law}(\bar X_t^{\leftarrow,\infty})$ and $p_t^{\infty,q_T} := \mathrm{law}(X_t^{\leftarrow,\infty,q_T})$. Likewise, denote the corresponding measures over path space by $Q_T^{\leftarrow,\infty}$ and P_T^{∞,q_T}. As it is convenient to think about a single stochastic process, for ease of notation we denote this by $(X_t^\infty)_{t\in[0,T]}$ so that under $Q_T^{\leftarrow,\infty}$ (resp. P_T^{∞,q_T}), $(X_t^\infty)_{t\in[0,T]}$ has the law of the modified reverse process (C.2) (resp. the modified SGM process (C.3)).

We first verify that these modified processes satisfy Novikov's condition.

Lemma 19. *It holds that*

$$\mathbb{E}_{Q_T^{\leftarrow,\infty}} \exp\Big(\sum_{k=0}^{N-1} \int_{kh}^{(k+1)h} \|\phi_R(X_{kh}^\infty)\,s^\infty_{T-kh}(X_{kh}^\infty) - \phi_R(X_t^\infty)\,\nabla \ln q_{T-t}(X_t^\infty)\|^2\,\mathrm{d}t\Big) < \infty\,.$$

Proof. For any x,

$$\begin{aligned}\|\phi_R(x)\,s^\infty_{T-kh}(x)\| &\le \sup_{x^*\in\mathsf{B}(0,R),\;k^*\in\{0,\dots,N-1\}} \|s^\infty_{T-k^*h}(x^*)\| =: A < \infty\,, \\ \|\phi_R(x)\,\nabla \ln q_{T-t}(x)\| &\le \sup_{x^*\in\mathsf{B}(0,R),\;t^*\in[0,T]} \|\nabla \ln q_{T-t^*}(x^*)\| =: B < \infty\,.\end{aligned}$$

So the quantity in the lemma is at most $\exp(2T\,(A^2+B^2)) < \infty$ as claimed. $\square$

We can thus apply Girsanov's theorem (Theorem 6), with the choices $P_T = P_T^{\infty,q_T}$, $Q_T = Q_T^{\leftarrow,\infty}$, $\tilde b_t = \phi_R(X_t^\infty)\,\{X_t^\infty + 2\,s_{T-kh}^\infty(X_{kh}^\infty)\}$, $b_t = \phi_R(X_t^\infty)\,\{X_t^\infty + 2\,\nabla \ln q_{T-t}(X_t^\infty)\}$, and $\sigma_t = \sqrt{2}\,I_d$. It yields the following corollary.

Corollary 20. *It holds that*

$$\begin{aligned}\mathsf{KL}(Q_T^{\leftarrow,\infty} \parallel P_T^{\infty,q_T}) &= \mathbb{E}_{Q_T^{\leftarrow,\infty}} \ln \frac{\mathrm{d}Q_T^{\leftarrow,\infty}}{\mathrm{d}P_T^{\infty,q_T}} \\ &= \sum_{k=0}^{N-1} \mathbb{E}_{Q_T^{\leftarrow,\infty}} \int_{kh}^{(k+1)h} \|\phi_R(X_{kh}^\infty)\, s_{T-kh}^\infty(X_{kh}^\infty) - \phi_R(X_t^\infty)\,\nabla \ln q_{T-t}(X_t^\infty)\|^2\,\mathrm{d}t\,.\end{aligned}$$

Our analysis for the quantity on the right-hand side of Corollary 20 proceeds along similar lines to the argument in Section 5.2, though the moment and movement bounds require some additional work as we must work with $\bar X_t^{\leftarrow,\infty}$ instead of $\bar X_t^{\leftarrow}$, the former of which is not the reverse of the OU process.

Theorem 21 (discretization error for modified process)**.** *Suppose that Assumptions 1, 2, and 3 hold. For $R > 0$, let $Q_T^{\leftarrow,\infty}$ and P_T^{∞,q_T} denote the measures on path space corresponding to the modified processes* (C.2) *and* (C.3) *respectively. Assume that $L \ge 1$ and $h \lesssim 1/L$. Then,*

$$\limsup_{R\to\infty} \mathsf{KL}(Q_T^{\leftarrow,\infty} \parallel P_T^{\infty,q_T}) \lesssim (\varepsilon_{\text{score}}^2 + L^2 dh + L^2 \mathfrak{m}_2^2 h^2)\,T\,.$$

Proof. First, as before, we can decompose

$$\begin{aligned}&\mathbb{E}_{Q_T^{\leftarrow,\infty}}[\|\phi_R(X_{kh}^\infty)\, s_{T-kh}^\infty(X_{kh}^\infty) - \phi_R(X_t^\infty)\,\nabla \ln q_{T-t}(X_t^\infty)\|^2] \\ &\quad\lesssim \mathbb{E}_{Q_T^{\leftarrow,\infty}}[\|\phi_R(X_{kh}^\infty)\,\{s_{T-kh}^\infty(X_{kh}^\infty) - \nabla \ln q_{T-kh}(X_{kh}^\infty)\}\|^2] \\ &\qquad\quad + \mathbb{E}_{Q_T^{\leftarrow,\infty}}[\|\phi_R(X_{kh}^\infty)\,\{\nabla \ln q_{T-kh}(X_{kh}^\infty) - \nabla \ln q_{T-t}(X_{kh}^\infty)\}\|^2] \\ &\qquad\quad + \mathbb{E}_{Q_T^{\leftarrow,\infty}}[\|\phi_R(X_{kh}^\infty)\,\nabla \ln q_{T-t}(X_{kh}^\infty) - \phi_R(X_t^\infty)\,\nabla \ln q_{T-t}(X_t^\infty)\|^2] \\ &\quad\lesssim \varepsilon_{\text{score}}^2 + o(1) + \mathbb{E}_{Q_T^{\leftarrow,\infty}}\Big[\Big\|\phi_R(X_{kh}^\infty)\,\nabla \ln \frac{q_{T-kh}}{q_{T-t}}(X_{kh}^\infty)\Big\|^2\Big] + L'^2\,\mathbb{E}_{Q_T^{\leftarrow,\infty}}[\|X_{kh}^\infty - X_t^\infty\|^2]\,,\end{aligned}$$

as $R \to \infty$, where in the last step we used the fact that $\phi_R(X_{kh}^\infty) \in [0,1]$ together with Lemma 18 applied to the vector field $v = \nabla \ln q_{T-t}$ (L' is the quantity defined therein).

Because (5.3) and (5.4) hold pointwise, we can bound the second term above by

$$\begin{aligned}\Big\|\phi_R(X_{kh}^\infty)\,\nabla \ln \frac{q_{T-kh}}{q_{T-t}}(X_{kh}^\infty)\Big\|^2 &\lesssim \phi_R(X_{kh}^\infty)^2\,\big(L^2 dh + L^2h^2\,\|X_{kh}^\infty\|^2 + L^2h^2\,\|\nabla \ln q_{T-t}(X_{kh}^\infty)\|^2\big) \\ &\lesssim L^2 dh + \phi_R(X_{kh}^\infty)^2\,L^2h^2\,\|X_{kh}^\infty\|^2 + \phi_R(X_t^\infty)^2\,L^2h^2\,\|\nabla \ln q_{T-t}(X_t^\infty)\|^2 + L'^2\,\|X_{kh}^\infty - X_t^\infty\|^2\,,\end{aligned}$$

where in the last line we applied Lemma 18 again. Hence,

$$\begin{aligned}&\mathbb{E}_{Q_T^{\leftarrow,\infty}}[\|\phi_R(X_{kh}^\infty)\, s_{T-kh}^\infty(X_{kh}^\infty) - \phi_R(X_t^\infty)\,\nabla \ln q_{T-t}(X_t^\infty)\|^2] \\ &\quad\lesssim \varepsilon_{\text{score}}^2 + o(1) + L^2 dh + \mathbb{E}_{Q_T^{\leftarrow,\infty}}[\phi_R(X_{kh}^\infty)^2\,L^2h^2\,\|X_{kh}^\infty\|^2] \\ &\qquad\quad + \mathbb{E}_{Q_T^{\leftarrow,\infty}}[\phi_R(X_t^\infty)^2\,L^2h^2\,\|\nabla \ln q_{T-t}(X_t^\infty)\|^2] + L'^2\,\mathbb{E}_{Q_T^{\leftarrow,\infty}}[\|X_{kh}^\infty - X_t^\infty\|^2]\,.\end{aligned}$$

By taking $R \to \infty$ above, noting that $\limsup_{R\to\infty} L' \lesssim L$, and applying Lemmas 23, 24, 25, and 26, we get

$$\begin{aligned}&\limsup_{R\to\infty} \mathbb{E}_{Q_T^{\leftarrow,\infty}}[\|\phi_R(X_{kh}^\infty)\, s_{T-kh}^\infty(X_{kh}^\infty) - \phi_R(X_t^\infty)\,\nabla \ln q_{T-t}(X_t^\infty)\|^2] \\ &\quad\lesssim \varepsilon_{\text{score}}^2 + L^2 dh + L^2h^2\,\mathbb{E}_{Q_T^{\leftarrow}}[\|X_t\|^2] + L^2h^2\,\mathbb{E}_{Q_T^{\leftarrow}}[\|\nabla \ln q_{T-t}(X_t)\|^2] + L^2\,(\mathfrak{m}_2^2 h^2 + dh)\,.\end{aligned}$$

We can bound $\mathbb{E}_{Q_T^{\leftarrow}}[\|X_t\|^2] \le d \vee \mathfrak{m}_2^2$ and $\mathbb{E}_{Q_T^{\leftarrow}}[\|\nabla \ln q_{T-t}(X_t)\|^2] \le Ld$ via the moment bounds of Lemma 9.

Putting everything together, we get

$$\limsup_{R\to\infty} \mathbb{E}[\|\phi_R(X_{kh}^\infty)\, s_{T-kh}^\infty(X_{kh}^\infty) - \phi_R(X_t^\infty)\, \nabla \ln q_{T-t}(X_t^\infty)\|^2] \lesssim \varepsilon_{\text{score}}^2 + L^2 dh + L^2 \mathfrak{m}_2^2 h^2\,.$$

The result now follows from Corollary 20. □

We are now ready to bound the discretization error for DDPM in total variation.

Proof. [Full proof of Theorem 8] By Theorem 21 and Pinsker's inequality, we have

$$\limsup_{R\to\infty} \mathsf{TV}(Q_T^{\leftarrow,\infty}, P_T^{\infty,q_T})^2 \lesssim (\varepsilon_{\text{score}}^2 + L^2 dh + L^2 \mathfrak{m}_2^2 h^2)\, T\,. \tag{C.4}$$

The plan now is to replace $Q_T^{\leftarrow,\infty}$ by $Q_T^{\leftarrow}$ and P_T^{∞,q_T} by $P_T^{q_T}$ via approximation arguments.

We begin with $Q_T^{\leftarrow,\infty}$. Consider a coupling of the laws of $(\bar X_t^{\leftarrow})_{t\in[0,T]}$ and $(\bar X_t^{\leftarrow,\infty})_{t\in[0,T]}$ under $Q_T^{\leftarrow}$ and $Q_T^{\leftarrow,\infty}$ such that if $\bar X_t^{\leftarrow} \in \mathsf{B}(0,R)$ for all $t\in[0,T]$, then $\bar X_t^{\leftarrow} = \bar X_t^{\leftarrow,\infty}$ for $t\in[0,T]$. Denote the complement of this event by $\mathcal{E}_0^R$. Then,

$$\mathsf{TV}(Q_T^{\leftarrow,\infty}, Q_T^{\leftarrow}) \le \mathbb{P}[\mathcal{E}_0^R] = o(1/R^2) \tag{C.5}$$

by Lemma 22 below.

Next, we would like to apply the same argument in order to handle $\mathsf{TV}(P_T^{\infty,q_T}, P_T^{q_T})$, but this is complicated by the fact that without assumptions on the score estimate s, it is not immediate that the process $X^{\leftarrow,q_T}$ under $P_T^{q_T}$ remains in the ball $\mathsf{B}(0,R)$ with high probability as $R\to\infty$. Indeed, since we do not assume any growth conditions on the score estimate s, we do not have any control over the moments of the process $X^{\leftarrow,q_T}$. To handle this complication, we introduce another auxiliary process $\tilde X^{\leftarrow,q_T}$ and use a change of measure argument similar in spirit to [LLT22, Theorem 4.1].

The auxiliary process $\tilde X^{\leftarrow,q_T}$ corresponds to the SGM algorithm with L^∞-accurate score and started from q_T: for $t\in[kh,(k+1)h]$,

$$\mathrm{d}\tilde X_t^{\leftarrow,q_T} = \{\tilde X_t^{\leftarrow,q_T} + 2\, s_{T-kh}^\infty(\tilde X_{kh}^{\leftarrow,q_T})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \qquad \tilde X_0^{\leftarrow,q_T} \sim q_T\,.$$

The difference between $\tilde X^{\leftarrow,q_T}$ and $X^{\leftarrow,\infty,q_T}$ is that in the definition of $\tilde X^{\leftarrow,q_T}$, we do not use the cutoff function ϕ_R; the difference between $\tilde X^{\leftarrow,q_T}$ and $X^{\leftarrow,q_T}$ is that in the definition of $\tilde X^{\leftarrow,q_T}$ we use the L^∞-accurate score s^∞. Let $\tilde P^{q_T}$ denote the corresponding path measure for $\tilde X^{\leftarrow,q_T}$. Consider a coupling of the laws of $(\tilde X_t^{\leftarrow,q_T})_{t\in[0,T]}$ and $(X_t^{\leftarrow,\infty,q_T})_{t\in[0,T]}$ under $\tilde P_T^{q_T}$ and P_T^{∞,q_T} such that if $\tilde X_t^{\leftarrow,q_T} \in \mathsf{B}(0,R)$ for all $t\in[0,T]$, then $X_t^{\leftarrow,\infty,q_T} = \tilde X_t^{\leftarrow,q_T}$ for $t\in[0,T]$. Denote the complement of this event by $\mathcal{E}_1^{R,\varepsilon_{\text{score},\infty}}$, so that

$$\mathsf{TV}(P_T^{\infty,q_T}, \tilde P_T^{q_T}) \le \mathbb{P}[\mathcal{E}_1^{R,\varepsilon_{\text{score},\infty}}]\,.$$

To bound this probability, we use a standard Grönwall argument. Since $\nabla \ln q_t$ is L-Lipschitz and $\|s_t^\infty - \nabla\ln q_t\| \le \varepsilon_{\text{score},\infty}$ for all $t\ge 0$, then for all $x\in\mathbb{R}^d$ and all $k=0,1,\dots,N-1$,

$$\|s_{kh}^\infty(x)\| \le \varepsilon_{\text{score},\infty} + \|\nabla \ln q_{kh}(x)\| \le \underbrace{\varepsilon_{\text{score},\infty} + \max_{k=0,1,\dots,N-1} \|\nabla\ln q_{kh}(0)\|}_{=:A(\varepsilon_{\text{score},\infty})} + L\,\|x\|\,.$$

Hence, for $u\in[kh,(k+1)h]$,

$$\begin{aligned}
\sup_{t\in[kh,u]} \|\tilde X_t^{\leftarrow,q_T}\| &= \sup_{t\in[kh,u]} \Bigl\|\tilde X_{kh}^{\leftarrow,q_T} + \int_{kh}^t \tilde X_s^{\leftarrow,q_T}\,\mathrm{d}s + (t-kh)\, s_{T-kh}^\infty(\tilde X_{kh}^{\leftarrow,q_T}) + \sqrt{2}\,(B_t - B_{kh})\Bigr\| \\
&\lesssim \|\tilde X_{kh}^{\leftarrow,q_T}\| + \int_{kh}^t \|\tilde X_s^{\leftarrow,q_T}\|\,\mathrm{d}s + h\,\|s_{T-kh}^\infty(\tilde X_{kh}^{\leftarrow,q_T})\| + \sup_{t\in[kh,(k+1)h]} \|B_t - B_{kh}\| \\
&\lesssim A(\varepsilon_{\text{score},\infty})\, h + (1+Lh)\,\|\tilde X_{kh}^{\leftarrow,q_T}\| + \sup_{t\in[kh,(k+1)h]} \|B_t - B_{kh}\| + h\int_{kh}^t \|\tilde X_s^{\leftarrow,q_T}\|\,\mathrm{d}s\,.
\end{aligned}$$

Assuming that $h \lesssim 1/L \le 1$, taking expectations and applying Grönwall's inequality yields

$$\mathbb{E} \sup_{t\in[kh,u]} \|\tilde{X}_t^{\leftarrow,q_T}\| \lesssim A(\varepsilon_{\text{score},\infty})\, h + \mathbb{E}\|\tilde{X}_{kh}^{\leftarrow,q_T}\| + \mathbb{E} \sup_{t\in[0,h]} \|B_t\| \,. \tag{C.6}$$

In particular,

$$\mathbb{E}\|\tilde{X}_{(k+1)h}^{\leftarrow,q_T}\| \lesssim A(\varepsilon_{\text{score},\infty})\, h + \mathbb{E}\|\tilde{X}_{kh}^{\leftarrow,q_T}\| + \mathbb{E} \sup_{t\in[0,h]} \|B_t\|$$

and iterating this yields the existence of $C > 0$ such that

$$\max_{k=0,1,\dots,N-1} \mathbb{E}\|\tilde{X}_{kh}^{\leftarrow,q_T}\| \lesssim \exp(CN) \left\{A(\varepsilon_{\text{score},\infty})\, h + \mathbb{E} \sup_{t\in[0,h]} \|B_t\| + \mathbb{E}_{q_T}\|\cdot\|\right\} < \infty \,.$$

Then, from (C.6),

$$\mathbb{E} \sup_{t\in[0,T]} \|\tilde{X}_t^{\leftarrow,q_T}\| \le \sum_{k=0}^{N-1} \mathbb{E} \sup_{t\in[kh,(k+1)h]} \|\tilde{X}_t^{\leftarrow,q_T}\| \le C(\varepsilon_{\text{score},\infty}) < \infty$$

for some constant $C(\varepsilon_{\text{score},\infty}) > 0$. From Markov's inequality, we deduce that

$$\mathbb{P}[\mathcal{E}_1^{R,\varepsilon_{\text{score},\infty}}] = \mathbb{P}\big[\sup_{t\in[0,T]} \|\tilde{X}_t^{\leftarrow,q_T}\| \ge R\big] \le \frac{C(\varepsilon_{\text{score},\infty})}{R} \,. \tag{C.7}$$

Combining together (C.4), (C.5), and (C.7),

$$\begin{aligned} \mathsf{TV}(Q_T^{\leftarrow}, \tilde{P}_T^{q_T}) &\le \limsup_{R\to\infty} \{\mathsf{TV}(Q_T^{\leftarrow,\infty}, P_T^{\infty,q_T}) + \mathsf{TV}(Q_T^{\leftarrow,\infty}, Q_T^{\leftarrow}) + \mathsf{TV}(P_T^{\infty,q_T}, \tilde{P}_T^{q_T})\} \\ &\lesssim (\varepsilon_{\text{score}} + L\sqrt{dh} + L\mathfrak{m}_2 h)\, \sqrt{T} \,. \end{aligned} \tag{C.8}$$

Next, consider a coupling of the processes $(X_t^{\leftarrow,q_T})_{t\in[0,T]}$ and $(\tilde{X}_t^{\leftarrow,q_T})_{t\in[0,T]}$ under $P_T^{q_T}$ and $\tilde{P}_T^{q_T}$ such that if $\tilde{X}_{kh}^{\leftarrow,q_T} \notin B_{T-kh}$ for $k = 0, 1, \dots, N-1$, then $X_t^{\leftarrow,q_T} = \tilde{X}_t^{\leftarrow,q_T}$ for all $t \in [0,T]$. Call the complement of this event $\mathcal{E}_2^{\varepsilon_{\text{score},\infty}}$. Hence,

$$\mathsf{TV}(\tilde{P}_T^{q_T}, P_T^{q_T}) \le \mathbb{P}[\mathcal{E}_2^{\varepsilon_{\text{score},\infty}}] \,. \tag{C.9}$$

By Chebyshev's inequality and the union bound, we have

$$\mathbb{P}[\bar{X}_{kh}^{\leftarrow} \in B_{T-kh} \text{ for some } k = 0, 1, \dots, N-1] \le \frac{N\varepsilon_{\text{score}}^2}{\varepsilon_{\text{score},\infty}^2} \,.$$

Using (C.8),

$$\begin{aligned} \mathbb{P}[\mathcal{E}_2^{\varepsilon_{\text{score},\infty}}] &= \mathbb{P}[\tilde{X}_{kh}^{\leftarrow,q_T} \in B_{T-kh} \text{ for some } k = 0, 1, \dots, N-1] \\ &\lesssim \frac{N\varepsilon_{\text{score}}^2}{\varepsilon_{\text{score},\infty}^2} + (\varepsilon_{\text{score}} + L\sqrt{dh} + L\mathfrak{m}_2 h)\, \sqrt{T} \,. \end{aligned} \tag{C.10}$$

By (C.8), (C.9) and (C.10),

$$\mathsf{TV}(Q_T^{\leftarrow}, P_T^{q_T}) \le \mathsf{TV}(Q_T^{\leftarrow}, \tilde{P}_T^{q_T}) + \mathsf{TV}(\tilde{P}_T^{q_T}, P_T^{q_T}) \lesssim \frac{N\varepsilon_{\text{score}}^2}{\varepsilon_{\text{score},\infty}^2} + (\varepsilon_{\text{score}} + L\sqrt{dh} + L\mathfrak{m}_2 h)\, \sqrt{T} \,.$$

Letting $\varepsilon_{\text{score},\infty} \to \infty$ we obtain the claimed result. $\square$

C.2 Auxiliary lemmas

In this section, we prove some auxiliary lemmas which were used in the proof of Theorem 21 above.

Lemma 22 (escape probability for reverse process)**.** *Under Assumption 2,*

$$\mathbb{P}\{\bar X_t^{\leftarrow} \in \mathsf{B}(0,R) \text{ for all } t \in [0,T]\} \ge 1 - o(1/R^2)\,.$$

Proof. As $\bar X^{\leftarrow}$ is the reverse of the OU process, we can equivalently bound $\mathbb{P}\{\bar X_t \in \mathsf{B}(0,R)$ for all $t \in [0,T]\}$. From the explicit solution of the OU process, $\bar X_t = \exp(-t)\,\bar X_0 + \sqrt{2}\int_0^t \exp(-(t-s))\,\mathrm{d}B_s$. Let $\tilde T \coloneqq \int_0^t \exp(-2\,(t-s))\,\mathrm{d}s$. By the representation of stochastic integrals as time changes of Brownian motion (see, e.g., [Ste01, Theorem 12.4]), it follows that $\sup_{t\in[0,T]} \|\int_0^t \exp(-(t-s))\,\mathrm{d}B_s\| = \sup_{t\in[0,\tilde T]} \|\tilde B_t\|$, where $\tilde B$ is another standard Brownian motion. Then, we can bound

$$\sup_{t\in[0,T]} \|\bar X_t\| \le \|\bar X_0\| + \sqrt{2}\sup_{t\in[0,\tilde T]} \|\tilde B_t\|\,.$$

By Assumption 2, we have $\mathbb{P}\{\|\bar X_0\| > R/2\} \le \mathbb{E}[\|\bar X_0\|^{2+\eta}]/(R/2)^{2+\eta} = o(1/R^2)$. On the other hand, by [Che+21b, Lemma 23] applied to $\lambda = 1/(4\tilde T)$,

$$\mathbb{P}\Bigl\{\sup_{t\in[0,\tilde T]} \|\tilde B_t\| > \frac{R}{2\sqrt{2}}\Bigr\} \le 3^d \exp\bigl(-\frac{R^2}{32\tilde T}\bigr) = o(1/R^2)\,.$$

The lemma follows from a union bound over the two events. □

Lemma 23 (effect of truncation on the score estimation error)**.** *We have*

$$\limsup_{R\to\infty} \mathbb{E}_{Q_T^{\leftarrow,\infty}}[\|\phi_R(X_{kh}^\infty)\,\{s_{T-kh}^\infty(X_{kh}^\infty) - \nabla \ln q_{T-kh}(X_{kh}^\infty)\}\|^2] \le \varepsilon_{\rm score}^2\,.$$

Proof. Consider a coupling of the laws of $(X_t)_{t\in[0,T]}$ and $(X_t^\infty)_{t\in[0,T]}$ such that if $X_{t'} \in \mathsf{B}(0,R)$ for all $0 \le t' < t$, then $X_t^\infty = X_t$. Denote the complement of this event by $\mathcal{E}_t$. Then we have

$$\begin{aligned}
&\mathbb{E}[\|\phi_R(X_{kh}^\infty)\,\{s_{T-kh}^\infty(X_{kh}^\infty) - \nabla \ln q_{T-kh}(X_{kh}^\infty)\}\|^2] \\
&\qquad \le \mathbb{E}[\|s_{T-kh}^\infty(X_{kh}) - \nabla \ln q_{T-kh}(X_{kh})\|^2] + \mathbb{E}[\|\phi_R(X_{kh}^\infty)\,\{s_{T-kh}^\infty(X_{kh}^\infty) - \nabla \ln q_{T-kh}(X_{kh}^\infty)\}\|^2\,\mathbb{1}_{\mathcal{E}_t}]\,.
\end{aligned}$$

For the first term, by the definition of s^∞ in (C.1),

$$\begin{aligned}
\mathbb{E}[\|s_{T-kh}^\infty(X_{kh}) - \nabla \ln q_{T-kh}(X_{kh})\|^2] &= \mathbb{E}[\|\{s_{T-kh}(X_{kh}) - \nabla \ln q_{T-kh}(X_{kh})\}\,\mathbb{1}_{B_{T-kh}^{\rm c}}\|^2] \\
&\le \mathbb{E}[\|s_{T-kh}(X_{kh}) - \nabla \ln q_{T-kh}(X_{kh})\|^2] \le \varepsilon_{\rm score}^2\,.
\end{aligned}$$

For the second term, by Lemma 22,

$$\mathbb{E}[\|\phi_R(X_{kh}^\infty)\,\{s_{T-kh}^\infty(X_{kh}^\infty) - \nabla \ln q_{T-kh}(X_{kh}^\infty)\}\|^2\,\mathbb{1}_{\mathcal{E}_t}] \le \varepsilon_{{\rm score},\infty}^2\,\mathbb{P}[\mathcal{E}_t] = o\bigl(\frac{\varepsilon_{{\rm score},\infty}^2}{R^2}\bigr)\,.$$

The result follows by taking $R \to \infty$. □

Lemma 24 (effect of truncation on the moment bound)**.** *Under Assumption 2, we have*

$$\limsup_{R\to\infty} \mathbb{E}_{Q_T^{\leftarrow,\infty}}[\phi_R(X_t^\infty)^2\,\|X_t^\infty\|^2] \le \mathbb{E}_{Q_T^{\leftarrow}}[\|X_t\|^2]\,.$$

Proof. Consider a coupling of the laws of $(X_t)_{t\in[0,T]}$ and $(X_t^\infty)_{t\in[0,T]}$ such that if $X_{t'} \in \mathsf{B}(0,R)$ for all $0 \le t' < t$, then $X_t^\infty = X_t$. Denote the complement of this event by $\mathcal{E}_t$. Then we have

$$\mathbb{E}[\phi_R(X_t^\infty)^2\,\|X_t^\infty\|^2] \le \mathbb{E}[\|X_t\|^2] + \mathbb{E}[\phi_R(X_t^\infty)^2\,\|X_t^\infty\|^2\,\mathbb{1}_{\mathcal{E}_t}]\,.$$

Because $\phi_R(X_t^\infty) = 0$ if $\|X_t^\infty\| \ge 2R$, we can bound the latter term via Hölder's inequality by

$$(2R)^2 \cdot \mathbb{P}[\mathcal{E}_t] = o(1)\,,$$

from Lemma 22. As this tends to 0 as $R \to \infty$, the lemma follows. □

Lemma 25 (effect of truncation on the score function bound)**.** *Under Assumptions 1 and 2,*

$$\limsup_{R\to\infty} \mathbb{E}_{Q_T^{\leftarrow,\infty}}[\phi_R(X_t^\infty)^2 \, \|\nabla \ln q_{T-t}(X_t^\infty)\|^2] \le \mathbb{E}_{Q_T^{\leftarrow}}[\|\nabla \ln q_{T-t}(X_t)\|^2]$$

for any $t \in [0,T]$.

Proof. Consider the same coupling of the laws of $(X_t)_{t\in[0,T]}$ and $(X_t^\infty)_{t\in[0,T]}$ from the proof of Lemma 24, and let $\mathcal{E}_t$ again denote the event that $X_{t'} \in \mathsf{B}(0,R)$ for all $0 \le t' < t$. Then we have

$$\mathbb{E}[\phi_R(X_t^\infty)^2 \, \|\nabla \ln q_{T-t}(X_t^\infty)\|^2] \le \mathbb{E}[\|\nabla \ln q_{T-t}(X_t)\|^2] + \mathbb{E}[\phi_R(X_t^\infty)^2 \, \|\nabla \ln q_{T-t}(X_t^\infty)\|^2 \, \mathbb{1}_{\mathcal{E}_t}]\,.$$

Because $\phi_R(X_t^\infty) = 0$ if $\|X_t^\infty\| \ge 2R$, we can bound the latter term via Hölder's inequality by

$$\sup_{x\in\mathsf{B}(0,2R)} \|\nabla \ln q_{T-t}(x)\|^2 \cdot \mathbb{P}[\mathcal{E}_t]\,.$$

By Assumption 1, if $\|x\| \le 2R$, then $\|\nabla \ln q_{T-t}(x)\|^2 \lesssim \|\nabla \ln q_{T-t}(0)\|^2 + L^2R^2$, so the above quantity is bounded via Lemma 22 by $o(1)$ as $R\to\infty$. □

Lemma 26 (movement bound for modified process)**.** *Suppose that Assumptions 1 and 2 hold. Then for* $0 \le s < t$ *satisfying* $t - s \le h$*, if* $h \lesssim 1/L$*, then*

$$\limsup_{R\to\infty} \mathbb{E}_{Q_T^{\leftarrow,\infty}}[\|X_t^\infty - X_s^\infty\|^2] \lesssim \mathfrak{m}_2^2 h^2 + dh\,.$$

Proof. In the proof, we drop the subscript $Q_T^{\leftarrow,\infty}$. We can write

$$\begin{aligned}\mathbb{E}[\|X_t^\infty - X_s^\infty\|^2] &= \mathbb{E}\Big[\Big\|\int_s^t \phi_R(X_r^\infty)\,\{X_r^\infty + 2\,\nabla \ln q_{T-r}(X_r^\infty)\}\,\mathrm{d}r + \sqrt{2}\,(B_t - B_s)\Big\|^2\Big] \\ &\lesssim h\int_s^t \mathbb{E}[\phi_R(X_r^\infty)^2\,\|X_r^\infty\|^2]\,\mathrm{d}r + h\int_s^t \mathbb{E}[\phi_R(X_r^\infty)^2\,\|\nabla \ln q_{T-r}(X_r^\infty)\|^2]\,\mathrm{d}r + dh\,.\end{aligned}$$

Taking the lim sup on both sides as $R\to\infty$, we can apply Lemmas 24 and 25 to get

$$\limsup_{R\to\infty} \mathbb{E}[\|X_t^\infty - X_s^\infty\|^2] \lesssim h\int_s^t \mathbb{E}[\|X_r^\infty\|^2]\,\mathrm{d}r + h\int_s^t \mathbb{E}[\|\nabla \ln q_{T-r}(X_r^\infty)\|^2]\,\mathrm{d}r + dh\,.$$

The claimed bound then follows from the moment and score function bounds of Lemma 9. □

References

[BGL14] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and geometry of Markov diffusion operators*. Vol. 348. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer, Cham, 2014, pp. xx+552.

[BMR22] A. Block, Y. Mroueh, and A. Rakhlin. "Generative modeling with denoising auto-encoders and Langevin sampling". In: *arXiv e-prints*, arXiv:2002.00107 (2022).

[Bru+21] J. Bruna, O. Regev, M. J. Song, and Y. Tang. "Continuous LWE". In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. 2021, pp. 694–707.

[Cao+22] H. Cao et al. "A survey on generative diffusion model". In: *arXiv e-prints*, arXiv:2209.02646 (2022).

[Cat+22] P. Cattiaux, G. Conforti, I. Gentil, and C. Léonard. "Time reversal of diffusion processes under a finite entropy condition". Sept. 2022.

[Che+18] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. "Underdamped Langevin MCMC: a non-asymptotic analysis". In: *Proceedings of the 31st Conference on Learning Theory*. Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 300–323.

[Che+21a] S. Chewi, M. A. Erdogdu, M. B. Li, R. Shen, and M. Zhang. "Analysis of Langevin Monte Carlo from Poincaré to log-Sobolev". In: *arXiv e-prints*, arXiv:2112.12662 (2021).

[Che+21b] S. Chewi et al. "Optimal dimension dependence of the Metropolis-adjusted Langevin algorithm". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by M. Belkin and S. Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, Aug. 2021, pp. 1260–1300.

[Che+22a] S. Chen, A. Gollakota, A. Klivans, and R. Meka. "Hardness of noise-free learning for two-hidden-layer neural networks". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 10709–10724.

[Che+22b] Y. Chen, S. Chewi, A. Salim, and A. Wibisono. "Improved analysis for a proximal algorithm for sampling". In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by P.-L. Loh and M. Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 2984–3014.

[Che22] S. Chewi. *Log-concave sampling*. Book draft available at https://chewisinho.github.io/. 2022.

[CLL22] S. Chen, J. Li, and Y. Li. "Learning (very) simple generative models is hard". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 35143–35155.

[Cro+22] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah. "Diffusion models in vision: a survey". In: *arXiv e-prints*, arXiv:2209.04747 (2022).

[De +21] V. De Bortoli, J. Thornton, J. Heng, and A. Doucet. "Diffusion Schrödinger bridge with applications to score-based generative modeling". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 17695–17709.

[De 22] V. De Bortoli. "Convergence of denoising diffusion models under the manifold hypothesis". In: *Transactions on Machine Learning Research* (2022).

[DKR22] A. S. Dalalyan, A. Karagulyan, and L. Riou-Durand. "Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets". In: *Journal of Machine Learning Research* 23.235 (2022), pp. 1–38.

[DKS17] I. Diakonikolas, D. M. Kane, and A. Stewart. "Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures". In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2017, pp. 73–84.

[DN21] P. Dhariwal and A. Nichol. "Diffusion models beat GANs on image synthesis". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 8780–8794.

[DR20] A. S. Dalalyan and L. Riou-Durand. "On sampling from a log-concave density using kinetic Langevin diffusions". In: *Bernoulli* 26.3 (2020), pp. 1956–1988.

[Dud02] R. M. Dudley. *Real analysis and probability*. Vol. 74. Cambridge Studies in Advanced Mathematics. Revised reprint of the 1989 original. Cambridge University Press, Cambridge, 2002, pp. x+555.

[DV21] A. Daniely and G. Vardi. "From local pseudorandom generators to hardness of learning". In: *Conference on Learning Theory*. PMLR. 2021, pp. 1358–1394.

[DVK22] T. Dockhorn, A. Vahdat, and K. Kreis. "Score-based generative modeling with critically-damped Langevin diffusion". In: *International Conference on Learning Representations*. 2022.

[Föl85] H. Föllmer. "An entropy approach to the time reversal of diffusion processes". In: *Stochastic differential systems (Marseille-Luminy, 1984)*. Vol. 69. Lect. Notes Control Inf. Sci. Springer, Berlin, 1985, pp. 156–163.

[GVV22] A. Gupte, N. Vafa, and V. Vaikuntanathan. “Continuous LWE is as hard as LWE & applications to learning Gaussian mixtures”. In: *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science—FOCS 2022*. IEEE Computer Soc., Los Alamitos, CA, [2022] ©2022, pp. 1162–1173.

[HJA20] J. Ho, A. Jain, and P. Abbeel. “Denoising diffusion probabilistic models”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851.

[Hyv05] A. Hyvärinen. “Estimation of non-normalized statistical models by score matching”. In: *J. Mach. Learn. Res.* 6 (2005), pp. 695–709.

[Kin+21] D. Kingma, T. Salimans, B. Poole, and J. Ho. “Variational diffusion models”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 21696–21707.

[Le 16] J.-F. Le Gall. *Brownian motion, martingales, and stochastic calculus*. French. Vol. 274. Graduate Texts in Mathematics. Springer, [Cham], 2016, pp. xiii+273.

[Liu+22] X. Liu, L. Wu, M. Ye, and Q. Liu. “Let us build bridges: understanding and extending diffusion generative models”. In: *arXiv preprint arXiv:2208.14699* (2022).

[LLT22] H. Lee, J. Lu, and Y. Tan. “Convergence for score-based generative modeling with polynomial complexity”. In: *Advances in Neural Information Processing Systems*. Ed. by A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho. 2022.

[LLT23] H. Lee, J. Lu, and Y. Tan. “Convergence of score-based generative modeling for general data distributions”. In: *Proceedings of the 34th International Conference on Algorithmic Learning Theory*. Ed. by S. Agrawal and F. Orabona. Vol. 201. Proceedings of Machine Learning Research. PMLR, Feb. 2023, pp. 946–985.

[LST21] Y. T. Lee, R. Shen, and K. Tian. “Structured logconcave sampling with a restricted Gaussian oracle”. In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by M. Belkin and S. Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, Aug. 2021, pp. 2993–3050.

[Ma+21] Y.-A. Ma et al. “Is there an analog of Nesterov acceleration for gradient-based MCMC?” In: *Bernoulli* 27.3 (2021), pp. 1942–1992.

[MS22] D. Mikulincer and Y. Shenfeld. “On the Lipschitz properties of transportation along heat flows”. In: *arXiv preprint arXiv:2201.01382* (2022).

[OV01] F. Otto and C. Villani. “Comment on: “Hypercontractivity of Hamilton–Jacobi equations”, by S. G. Bobkov, I. Gentil and M. Ledoux”. In: *J. Math. Pures Appl. (9)* 80.7 (2001), pp. 697–700.

[Pid22] J. Pidstrigach. “Score-based generative models detect manifolds”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 35852–35865.

[Ram+22] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. “Hierarchical text-conditional image generation with CLIP latents”. In: *arXiv preprint arXiv:2204.06125* (2022).

[SE19] Y. Song and S. Ermon. “Generative modeling by estimating gradients of the data distribution”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019.

[SL19] R. Shen and Y. T. Lee. “The randomized midpoint method for log-concave sampling”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019.

[Soh+15] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. “Deep unsupervised learning using nonequilibrium thermodynamics”. In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by F. Bach and D. Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 2256–2265.

[Son+21a] Y. Song, C. Durkan, I. Murray, and S. Ermon. “Maximum likelihood training of score-based diffusion models”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 1415–1428.

[Son+21b] Y. Song et al. "Score-based generative modeling through stochastic differential equations". In: *International Conference on Learning Representations*. 2021.

[Ste01] J. M. Steele. *Stochastic calculus and financial applications*. Vol. 45. Applications of Mathematics (New York). Springer-Verlag, New York, 2001, pp. x+300.

[Vil09] C. Villani. "Hypocoercivity". In: *Mem. Amer. Math. Soc.* 202.950 (2009), pp. iv+141.

[Vin11] P. Vincent. "A connection between score matching and denoising autoencoders". In: *Neural Comput.* 23.7 (2011), pp. 1661–1674.

[VKK21] A. Vahdat, K. Kreis, and J. Kautz. "Score-based generative modeling in latent space". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 11287–11302.

[VW19] S. Vempala and A. Wibisono. "Rapid convergence of the unadjusted Langevin algorithm: isoperimetry suffices". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8094–8106.

[Yan+22] L. Yang, Z. Zhang, S. Hong, W. Zhang, and B. Cui. "Diffusion models: a comprehensive survey of methods and applications". In: *arXiv e-prints*, arXiv:2209.00796 (2022).

[ZC23] Q. Zhang and Y. Chen. "Fast sampling of diffusion models with exponential integrator". In: *The Eleventh International Conference on Learning Representations*. 2023.