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

Anonymous Author(s)
Affiliation
Address
email

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 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.

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 explain in Section 2, there are several other sources of error which could cause the SGM to fail to converge. Indeed, despite a flurry of recent work [BMR20; De +21; De 22; LLT22; Pid22], prior analyses fall short of answering this question, for (at least) one of three main reasons:

Under review at the NeurIPS 2022 Workshop on Score-Based Methods. Do not distribute.

1. **Super-polynomial convergence.** The bounds obtained are not quantitative (e.g., [De +21; Pid22]), or scale exponentially in the dimension and other problem parameters [BMR20; 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 Appendix B for details), and there are empirical studies suggesting that in practice, the score estimate is not in fact L^∞-accurate (e.g., [ZC22]). Intuitively, this is because we cannot expect that the score estimate that 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.

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. 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 assumptions also cover data distributions satisfying the *manifold hypothesis*, i.e., distributions supported on a low-dimensional submanifold.

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.

Finally, in the Appendix, we study SGMs based on the critically damped Langevin diffusion (CLD) as proposed by [DVK22]. Contrary to popular wisdom, we give evidence that the use of CLDs does not improve the complexity of SGMs.

2 Approach

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 that we want to draw samples from. When q has a positive density, we write it as $q = \exp(-U)$, where $U : \mathbb{R}^d \to \mathbb{R}$ is the *potential*.

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 $(B_t)_{t\geq 0}$ is a standard Brownian motion in $\mathbb{R}^d$. This forward process can be implemented in closed form, and it transforms the data distribution q to a noisy distribution q_T at some terminal time T. Besides, q_T converges to the standard Gaussian γ^d exponentially fast in Kullback-Leibler divergence: $\mathsf{KL}(q_T \parallel \gamma^d) \leq \exp(-2T)\,\mathsf{KL}(q \parallel \gamma^d)$.

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. Here, the reverse process satisfies

$$\mathrm{d}\bar{X}_t^{\leftarrow} = \{\bar{X}_t^{\leftarrow} + 2\,\nabla \ln q_{T-t}(\bar{X}_t^{\leftarrow})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \qquad \bar{X}_0^{\leftarrow} \sim q_T\,, \tag{2.2}$$

where $(B_t)_{t\in[0,T]}$ is another Brownian motion. If this process is initialized at $\bar{X}_0^{\leftarrow} \sim q_T$, then $\bar{X}_T^{\leftarrow} \sim q$. However, in practice, the user cannot draw samples from q_T exactly. Instead, the user draws sample from $\gamma^d \simeq q_T$, and initializes the reverse process at γ^d. This way, a distribution γ_T^d would be obtained at time T, and using Pinsker's inequality and the data processing inequality:

$$\mathsf{TV}^2(q, \gamma_T^d) \lesssim \mathsf{KL}(q \parallel \gamma_T^d) \le \mathsf{KL}(q_T \parallel \gamma^d) \le \exp(-2T)\, \mathsf{KL}(q \parallel \gamma^d)\,.$$

This provides a guarantee for the closeness between the output γ_T^d of the "algorithm", and the data distribution q that we want to sample from.

Although the story above is simple, the situation is more complicated in practice: the reverse SDE cannot be implemented exactly, because of the two following reasons.

1. The vector $\nabla \ln q_t$, called the *score function* for q_t, is not explicitly known and is in practice learned on the basis of samples (which are obtained by running the forward process initialized at the samples from q, i.e., the dataset).
2. Even if the score function were known for every t, one cannot implement the reverse process in general because its solution does not admit a closed form when initialized from γ^d. In practice, the reverse process is implemented via time discretization using an Euler scheme.

Implementing the reverse process from γ^d with Euler time discretization *and* an L^2-accurate score estimate is what we call the DDPM algorithm (see Appendix B for details). Below, we state our results for the DDPM algorithm, which handle the two sources of error mentioned above. We first consider the case where the score functions of the forward process are Lipschitz, and then we consider the case where q is compactly supported (without assuming Lipschitzness of the score functions).

3 Results

We now state our assumptions and our main results. We make the following assumptions on the data distribution q.

Assumption 1 (Lipschitz score)**.** *For all $t \ge 0$, 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 [BMR20; LLT22]. However, unlike [LLT22], we do not assume Lipschitzness of the score estimate. Moreover, unlike [BMR20; 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.

We also assume an L^2 bound on the score estimation error.

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

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

This is the same assumption as in [LLT22], and it is a natural and realistic assumption in light of the score matching objective often used in practice (see Appendix B).

Our main result for the DDPM algorithm is the following theorem.

Theorem 1 (DDPM)**.** *Suppose that Assumptions 1, 2, and 3 hold. Let p_T be the output of the DDPM algorithm 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*

$$\mathsf{TV}(p_T, q) \lesssim \underbrace{\sqrt{\mathsf{KL}(q \parallel \gamma^d)}\exp(-T)}_{\textit{convergence of forward process}} + \underbrace{(L\sqrt{dh} + L\mathfrak{m}_2 h)\sqrt{T}}_{\textit{discretization error}} + \underbrace{\frac{N^{1/3}\varepsilon_{\text{score}}^{2/3}}{T^{1/6}}}_{\textit{score estimation error}}\,.$$

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

$$\mathsf{TV}(p_T, q) \leq \widetilde{O}\Big(\varepsilon + \frac{L^{2/3} d^{1/3} \varepsilon_{\text{score}}^{2/3}}{\varepsilon^{2/3}}\Big), \qquad \text{for } N = \widetilde{\Theta}\Big(\frac{L^2 d}{\varepsilon^2}\Big).$$

In particular, in order to have $\mathsf{TV}(p_T, q) \leq \varepsilon$, it suffices to have score error

$$\varepsilon_{\text{score}} \leq \widetilde{O}\Big(\frac{\varepsilon^{5/2}}{L d^{1/2}}\Big).$$

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.

3.1 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, $\operatorname{supp} q \subseteq \mathsf{B}(0, R)$, an assumption which is satisfied by data distributions in practical settings.

For this setting, Theorem 1 does 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 1 has a term involving $\mathsf{KL}(q \parallel \gamma^d)$ which is infinite if q is not absolutely continuous w.r.t. γ^d, for instance under the manifold hypothesis. As pointed out by [De 22], in general we cannot obtain non-trivial guarantees for $\mathsf{TV}(p_T, q)$, because the output of the algorithm 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. We can apply our Theorem 1 to q_t instead of q for a well chosen and small value of t. This way, we get the Corollary 2. Note that taking q_t as the new target distribution 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).

Corollary 2 (compactly supported data)**.** *Suppose that q is supported on the ball of radius $R \geq 1$. 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 1 and*

$$N = \widetilde{\Theta}\Big(\frac{d^3 R^4 \left(R \vee \sqrt{d}\right)^4}{\varepsilon_{\mathrm{TV}}^2\, \varepsilon_{W_2}^8}\Big) \qquad \textit{and} \qquad \varepsilon_{\text{score}} \leq \widetilde{O}\Big(\frac{\varepsilon_{\mathrm{TV}}^{5/2}\, \varepsilon_{W_2}^4}{d^{3/2} R^2 \left(R \vee \sqrt{d}\right)^2}\Big).$$

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].

4 Conclusion

Our results show that, given an L^2-accurate score estimate, SGMs can sample from (essentially) any data distribution q with polynomial complexity: q can be highly non-log-concave (in which case sampling from q is usually intractable), or supported on a low dimensional submanifold. In particular, we answer the open question of [LLT22] regarding whether or not SGMs can sample from multimodal distributions.

In the context of neural networks, our results imply that so long as the neural network succeeds at the score 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.

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.

[BMR20] A. Block, Y. Mroueh, and A. Rakhlin. "Generative modeling with denoising auto-encoders and Langevin sampling". In: *arXiv e-prints*, arXiv:2002.00107 (2020).

[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+22] 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.

[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: *arXiv e-prints*, arXiv:2208.05314 (2022).

[DKR19] A. S. Dalalyan, A. Karagulyan, and L. Riou-Durand. "Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets". In: *arXiv e-prints*, arXiv:1906.08530 (June 2019).

[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.

[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.

[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.

[LLT22] H. Lee, J. Lu, and Y. Tan. "Convergence for score-based generative modeling with polynomial complexity". In: *arXiv e-prints*, arXiv:2206.06227 (2022).

[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: *arXiv e-prints*, arXiv:2206.01018 (2022).

[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).

[ZC22] Q. Zhang and Y. Chen. "Fast sampling of diffusion models with exponential integrator". In: *arXiv e-prints*, arXiv:2204.13902 (2022).

A Prior work

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

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 [BMR20; LLT22][1] instead analyze SGMs under the more realistic assumption of an L^2-accurate score estimate. However, the bounds of [BMR20] 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 3.1. Unfortunately, the bounds of [De 22] also scale exponentially in problem parameters such as the manifold diameter.

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+22], and the present work can be viewed as expanding upon the theory of [Che+22] using a different forward channel (the OU process).

B Further details for DDPM implementation

In this section, we provide further background on SGMs.

Reversing an SDE. 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 $(\sigma_t)_{t\geq 0}$ 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

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

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

$$b_t + b_{T-t}^{\leftarrow} = \sigma_t\sigma_t^{\mathsf{T}}\nabla \ln q_t\,, \qquad q_t := \mathrm{law}(\bar{X}_t)\,. \tag{B.1}$$

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 $\mathcal{F}$,

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

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 (B.2) is *equivalent* to the following problem:

$$\underset{s_t\in\mathcal{F}}{\text{minimize}} \quad \mathbb{E}\Big[\Big\|s_t(\bar{X}_t) + \frac{1}{\sqrt{1-\exp(-2t)}}\,Z_t\Big\|^2\Big]\,, \tag{B.3}$$

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 (B.2) and (B.3) share the same minimizers. We give a self-contained derivation below

[1] Unfortunately, the current analysis of [BMR20] contains a gap, as Theorem 27 therein is false: an LSI does not imply contraction in the Wasserstein metric for the Langevin diffusion.

for the sake of completeness. Unlike (B.2), however, the objective in (B.3) 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

$$\underset{s_t \in \mathcal{F}}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^{n} \left\| s_t(\bar{X}_t^{(i)}) + \frac{1}{\sqrt{1-\exp(-2t)}} Z_t^{(i)} \right\|^2 , \tag{B.4}$$

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

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

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

We remark that given the objective function (B.2), it is most natural to assume an $L^2(q_t)$ error bound $\mathbb{E}_{q_t}[\|s_t - \nabla \ln q_t\|^2] \le \varepsilon_{\text{score}}^2$ 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 [BMR20].

Derivation of the score matching objective. We present a self-contained derivation of the score matching objective (B.3) for the reader's convenience. See also [Hyv05; Vin11; SE19].

Recall that the problem is to solve

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

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 $X_t \coloneqq \exp(-t)\, X_0 + \sqrt{1-\exp(-2t)}\, Z_t$.

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 (B.4), 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 (2.2), 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

$$\mathrm{d}X_t^{\leftarrow} = \{X_t^{\leftarrow} + 2\, s_{T-kh}(X_{kh}^{\leftarrow})\}\, \mathrm{d}t + \sqrt{2}\, \mathrm{d}B_t\,, \qquad t \in [kh, (k+1)h]\,. \tag{B.5}$$

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 (2.2) 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.

C Technical overview

We now give a detailed technical overview for the proof for DDPM (Theorem 1). The proof for CLD (Theorem 10) 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 (2.2), 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 (2.2) 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+22] 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), but similarly to [LLT22] we first assume that the score estimate is L^∞-accurate:

$$\max_{k=0,1,\dots,N-1} \|s_{kh} - \nabla \ln q_{kh}\|_{\sup} \le \varepsilon_{\mathrm{score}}\,. \tag{C.1}$$

Let p_t^∞ denote the law of the algorithm at time t with L^∞-accurate score estimate. 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 $\partial_t \mathsf{d}(q_{T-t}, p_t^\infty)$, which is used to bound $\mathsf{d}(q_{T-(k+1)h}, p_{(k+1)h}^\infty)$ in terms of $\mathsf{d}(q_{T-kh}, p_{kh}^\infty)$ 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.

Finally, we must replace the assumption (C.1) of an L^∞-accurate score estimate with Assumption 3, for which we take inspiration from [LLT22]. For $k = 0, 1, \ldots, (N-1)h$, define the bad set

$$B_{kh} \coloneqq \{\|s_{kh} - \nabla \ln q_{kh}\| \geq \varepsilon_{\mathrm{score},\infty}\}\,.$$

If the algorithm never enters the bad sets, i.e., $X_{kh}^{\leftarrow} \notin B_{T-kh}$ for $k = 0, 1, \ldots, N-1$, then it is as if the algorithm is using an L^∞-accurate score estimate, and we can apply the Girsanov discretization analysis. Therefore, if we can control the error probability

$$\mathbb{P}(X_{kh}^{\leftarrow} \in B_{T-kh} \text{ for some } k = 0, 1, \ldots, N-1) \leq \cdots\,,$$

then by a coupling argument, we can bound the overall total variation error $\mathsf{TV}(p_T, q)$.

However, by Chebyshev's inequality, we have $q_{kh}(B_{kh}) \leq \varepsilon_{\mathrm{score}}^2 / \varepsilon_{\mathrm{score},\infty}^2 =: \delta$, and by a union bound,

$$\mathbb{P}(\bar{X}_{kh}^{\leftarrow} \in B_{T-kh} \text{ for some } k = 0, 1, \ldots, N-1) \leq N\delta\,.$$

In brief, the difficulty now is that our assumption allows us to control the probability of the bad event for the reverse process $\bar{X}^{\leftarrow}$, but we need to control the probability under the algorithm $X^{\leftarrow}$. The solution is to use a change-of-measure inequality which relates these two probabilities. In [LLT22], the authors use the inequality

$$\mathbb{P}(X_{kh}^{\leftarrow} \in B_{T-kh} \text{ for some } k = 0, 1, \ldots, N-1) \leq N\sqrt{\delta \max_{k=0,1,\ldots,N-1} \chi^2(p_{kh}^\infty \,\|\, q_{T-kh})}\,.$$

In our work, we refine this approach as follows. First, we note that because the Girsanov approach bounds the divergence between measures on path space, it yields for free a stronger result than the interpolation method. Namely, rather than bounding the divergence $\mathsf{d}(q_{T-kh}, p_{kh}^{\infty,q_T})$ (here, p_{kh}^{∞,q_T} denotes the law of the algorithm at time kh with L^∞-accurate score, but initialized at q_T), it instead bounds the divergence between the *joint distributions*

$$\mathsf{d}\big(\mathrm{law}(\bar{X}_{kh}^{\leftarrow} : k = 0, 1, \ldots, N),\ \mathrm{law}(X_{kh}^{\leftarrow,\infty,q_T} : k = 0, 1, \ldots, N)\big) \leq \varepsilon_{\mathrm{Girsanov}}\,.$$

In particular, suppose we take d to be any divergence that upper bounds the TV distance (e.g., $\mathsf{d} = \sqrt{\mathsf{KL}}$). Using a TV bound between the joint distributions, we show below in Section D.3 that

$$\mathbb{P}(X_{kh}^{\leftarrow} \in B_{T-kh} \text{ for some } k = 0, 1, \ldots, N-1) \leq \mathsf{TV}(q_T, \gamma^d) + \varepsilon_{\mathrm{Girsanov}} + N\delta\,. \tag{C.2}$$

These three terms correspond to the components of the bound in Theorem 1: the convergence of the forward process, the discretization error, and the probability of failure. For the third term, the parameter δ is balanced in the final bound to optimize a trade-off (a smaller choice of δ leads to a larger value for $\varepsilon_{\mathrm{score},\infty}$, which in turn increases the error $\varepsilon_{\mathrm{Girsanov}}$), ultimately leading to a bound in terms of the score estimation error $\varepsilon_{\mathrm{score}}$.

We conclude this section by stating and proving the elementary lemma that leads to (C.2), and providing a list of notation for the proofs. Then, in Section D, we give the Girsanov discretization argument and conclude the proof of Theorem 1.

Lemma 3. *Suppose that $(\bar{X}_{kh}^{\leftarrow})_{k\in\mathbb{N}}$, $(X_{kh}^{\leftarrow})_{k\in\mathbb{N}}$, and $(X_{kh}^{\leftarrow,\infty})_{k\in\mathbb{N}}$ are random processes defined on the same probability space. For each $k \in \mathbb{N}$, let $q_{kh}^{\leftarrow}$, p_{kh}, and p_{kh}^∞ denote the laws of $\bar{X}_{kh}^{\leftarrow}$, $X_{kh}^{\leftarrow}$, and $X_{kh}^{\leftarrow,\infty}$ respectively.*

Suppose that $(B_{kh})_{k\in\mathbb{N}}$ are measurable subsets of $\mathbb{R}^d$ and that the following hold.

1. *If* $X_{kh}^{\leftarrow,\infty} \notin B_{T-kh}$ *for* $k = 0, 1, \ldots, N-1$*, then* $X_{kh}^{\leftarrow} = X_{kh}^{\leftarrow,\infty}$ *for all* $k = 0, 1, \ldots, N$.
2. *It holds that*
$$\mathsf{TV}\big(\mathrm{law}(X_{kh}^{\leftarrow,\infty} : k = 0, 1, \ldots, N-1),\ \mathrm{law}(\bar{X}_{kh}^{\leftarrow} : k = 0, 1, \ldots, N-1)\big) \le \varepsilon\,.$$
3. $\mathbb{P}[\bar{X}_{kh}^{\leftarrow} \in B_{T-kh}] \le \delta$ *for all* $k = 0, 1, \ldots, N-1$.

Then, the following holds:
$$\mathsf{TV}(p_{Nh}, q_{Nh}^{\leftarrow}) \le 2\varepsilon + N\delta\,.$$

Proof. By triangle inequality,
$$\mathsf{TV}(p_{Nh}, q_{Nh}^{\leftarrow}) \le \mathsf{TV}(p_{Nh}, p_{Nh}^{\infty}) + \mathsf{TV}(p_{Nh}^{\infty}, q_{Nh}^{\leftarrow}) \le \mathsf{TV}(p_{Nh}, p_{Nh}^{\infty}) + \varepsilon\,,$$
where we used the data processing inequality. Then, by the union bound
$$\begin{aligned}\mathsf{TV}(p_{Nh}, p_{Nh}^{\infty}) &\le \mathbb{P}(X_{Nh}^{\leftarrow} \ne X_{Nh}^{\leftarrow,\infty}) \le \mathbb{P}(X_{kh}^{\leftarrow,\infty} \in B_{T-kh} \text{ for some } k = 0, 1, \ldots, N-1)\\ &\le \varepsilon + \mathbb{P}(\bar{X}_{kh}^{\leftarrow} \in B_{T-kh} \text{ for some } k = 0, 1, \ldots, N-1)\\ &\le \varepsilon + \sum_{k=0}^{N-1} \mathbb{P}(\bar{X}_{kh}^{\leftarrow} \in B_{kh}) \le \varepsilon + N\delta\,.\end{aligned}$$
This proves the lemma. □

Notation

L^∞-accurate score estimate. The bad set is defined to be
$$B_t := \{\|s_t - \nabla \ln q_t\| \ge \varepsilon_{\mathrm{score},\infty}\}\,,$$
where $\varepsilon_{\mathrm{score},\infty} > 0$ is a parameter to be chosen later. We define the L^∞-accurate score estimate to be
$$s_t^{\infty} := s_t\, \mathbb{1}_{B_t^c} + \nabla \ln q_t\, \mathbb{1}_{B_t}\,. \tag{C.3}$$
We note that $\|s_t^{\infty} - \nabla \ln q_t\| \le \varepsilon_{\mathrm{score},\infty}$.

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 (2.2) is denoted $(\bar{X}_t^{\leftarrow})_{t\in[0,T]}$, where $\bar{X}_t^{\leftarrow} := \bar{X}_{T-t} \sim q_{T-t}$.
- The SGM algorithm (B.5) is denoted $(X_t^{\leftarrow})_{t\in[0,T]}$, and $X_t^{\leftarrow} \sim p_t$. Recall that we initialize at $p_0 = \gamma^d$, the standard Gaussian measure.
- The SGM algorithm with L^∞-accurate score, denoted $(X_t^{\leftarrow,\infty})_{t\in[0,T]}$, is the same as (B.5) except that we use s_{T-kh}^{∞} defined in (C.3) instead of s_{T-kh} for $k = 0, 1, \ldots, N-1$. We write $X_t^{\leftarrow,\infty} \sim p_t^{\infty}$, and we initialize the process at $p_0^{\infty} = \gamma^d$.
- The process $(X_t^{\leftarrow,\infty,q_T})_{t\in[0,T]}$ is the same as $(X_t^{\leftarrow,\infty})_{t\in[0,T]}$, except that we initialize this process at q_T rather than at γ^d. We write $X_t^{\leftarrow,\infty,q_T} \sim p_t^{\infty,q_T}$.

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:

- $Q_T^{\leftarrow}$, under which $(X_t)_{t\in[0,T]}$ has the law of the reverse process (2.2); and
- P_T^{∞,q_T}, under which $(X_t)_{t\in[0,T]}$ has the law of the SGM algorithm with L^∞-accurate score initialized at q_T (corresponding to the process $(X_t^{\leftarrow,\infty,q_T})_{t\in[0,T]}$ 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.

D Proofs for DDPM

D.1 Preliminaries on Girsanov's theorem

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

Theorem 4 (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,

$$\mathbb{E}_{Q_T} \exp\Big(\frac{1}{2}\int_0^T \|\sigma_t^{-1}\,(\tilde{b}_t - b_t)\|^2 \,\mathrm{d}t\Big) < \infty \,, \tag{D.1}$$

we have that

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

We would like to apply Girsanov's theorem with $P_T = P_T^{\infty,q_T}$, $Q_T = Q_T^{\leftarrow}$, $\tilde{b}_t = X_t + 2\,s_{T-kh}^{\infty}(X_{kh})$ (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.[2] In Appendix G we give a workaround for this issue by defining a sequence of 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 work under the assumption that P_T^{∞,q_T} and $Q_T^{\leftarrow}$ satisfy Novikov's condition as it leads to a considerably simpler proof.

Corollary 5. *Assuming that P_T^{∞,q_T} and Q_T satisfy Novikov's condition (D.1), it holds that*

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

D.2 Girsanov discretization argument

We now apply Corollary 5 and bound the discretization error.

Theorem 6 (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 (2.2) and the SGM algorithm with L^∞-accurate score estimate initialized at q_T. Assume that $L \geq 1$ and $h \lesssim 1/L$. Then, it holds that*

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

We first give a proof under the assumption that P_T^{∞,q_T} and $Q_T^{\leftarrow}$ satisfy Novikov's condition (D.1). In Appendix G, we show how to lift this assumption and give a fully rigorous proof of this theorem.

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

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

$$\begin{aligned}
&\|s^{\infty}_{T-kh}(X_{kh}) - \nabla \ln q_{T-t}(X_t)\|^2 \\
&\quad \lesssim \|s^{\infty}_{T-kh}(X_{kh}) - \nabla \ln q_{T-kh}(X_{kh})\|^2 + \|\nabla \ln q_{T-kh}(X_{kh}) - \nabla \ln q_{T-t}(X_{kh})\|^2 \\
&\qquad + \|\nabla \ln q_{T-t}(X_{kh}) - \nabla \ln q_{T-t}(X_t)\|^2 \\
&\quad \lesssim \varepsilon^2_{\text{score},\infty} + \left\|\nabla \ln \frac{q_{T-kh}}{q_{T-t}}(X_{kh})\right\|^2 + L^2 \|X_{kh} - X_t\|^2 . \qquad \text{(D.2)}
\end{aligned}$$

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 14 that we prove in Section F.3) with $\alpha = \exp(t-kh) = 1 + O(h)$ and $\sigma^2 = 1 - \exp(-2\,(t-kh)) = O(h)$ to obtain

$$\begin{aligned}
\left\|\nabla \ln \frac{q_{T-kh}}{q_{T-t}}(X_{kh})\right\|^2 &\lesssim L^2 dh + L^2 h^2 \|X_{kh}\|^2 + (1 + L^2)\, h^2 \|\nabla \ln q_{T-t}(X_{kh})\|^2 \\
&\lesssim L^2 dh + L^2 h^2 \|X_{kh}\|^2 + L^2 h^2 \|\nabla \ln q_{T-t}(X_{kh})\|^2 \qquad \text{(D.3)}
\end{aligned}$$

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

For the last term,

$$\begin{aligned}
\|\nabla \ln q_{T-t}(X_{kh})\|^2 &\lesssim \|\nabla \ln q_{T-t}(X_t)\|^2 + \|\nabla \ln q_{T-t}(X_{kh}) - \nabla \ln q_{T-t}(X_t)\|^2 \\
&\lesssim \|\nabla \ln q_{T-t}(X_t)\|^2 + L^2 \|X_{kh} - X_t\|^2 , \qquad \text{(D.4)}
\end{aligned}$$

where the second term above is absorbed into the third term of the decomposition (D.2). Hence,

$$\begin{aligned}
\|s^{\infty}_{T-kh}(X_{kh}) - \nabla \ln q_{T-t}(X_t)\|^2 &\lesssim \varepsilon^2_{\text{score},\infty} + L^2 dh + L^2 h^2 \|X_{kh}\|^2 \\
&\quad + L^2 h^2 \|\nabla \ln q_{T-t}(X_t)\|^2 + L^2 \|X_{kh} - X_t\|^2 .
\end{aligned}$$

Using the fact that under $Q^{\leftarrow}_T$, 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 8 and the movement bound in Lemma 9:

$$\begin{aligned}
&\mathbb{E}_{Q^{\leftarrow}_T}[\|s^{\infty}_{T-kh}(X_{kh}) - \nabla \ln q_{T-t}(X_t)\|^2] \\
&\quad \lesssim \varepsilon^2_{\text{score},\infty} + L^2 dh + L^2 h^2 (d + \mathfrak{m}_2^2) + L^3 dh^2 + L^2 (\mathfrak{m}_2^2 h^2 + dh) \\
&\quad \lesssim \varepsilon^2_{\text{score},\infty} + L^2 dh + L^2 \mathfrak{m}_2^2 h^2 .
\end{aligned}$$

The result now follows from Corollary 5 and Pinsker's inequality. □

D.3 Proof of Theorem 1

We can now conclude our main result, namely a convergence guarantee when we only have an L^2-accurate score estimate. To prove this, we use the following truncation lemma.

Proof. [Proof of Theorem 1] We recall the notation from Section C. By the data processing inequality, Theorem 6 implies

$$\begin{aligned}
&\mathsf{TV}\big(\mathrm{law}(X^{\leftarrow,\infty,q_T}_{kh} : k = 0, 1, \ldots, N-1),\ \mathrm{law}(\bar{X}^{\leftarrow}_{kh} : k = 0, 1, \ldots, N-1)\big) \\
&\quad \lesssim (\varepsilon_{\text{score},\infty} + L\sqrt{dh} + L\mathfrak{m}_2 h)\sqrt{T} .
\end{aligned}$$

By the data processing inequality and Pinsker's inequality,

$$\begin{aligned}
&\mathsf{TV}\big(\mathrm{law}(X^{\leftarrow,\infty,q_T}_{kh} : k = 0, 1, \ldots, N-1),\ \mathrm{law}(X^{\leftarrow,\infty}_{kh} : k = 0, 1, \ldots, N-1)\big) \\
&\quad \leq \mathsf{TV}(q_T, \gamma^d) \leq \sqrt{\mathsf{KL}(q_T \,\|\, \gamma^d)} \leq \sqrt{\mathsf{KL}(q \,\|\, \gamma^d)}\exp(-T) ,
\end{aligned}$$

where we used the convergence of the OU process in KL divergence (see, e.g., [BGL14, Theorem 5.2.1]). Hence, we can apply Lemma 3. Here, we take the bad set $B_t \coloneqq \{\|s_t - \nabla \ln q_t\| \geq \varepsilon_{\text{score},\infty}\}$, so by Chebyshev's inequality and Assumption 3, we can take $\delta = (\varepsilon_{\text{score}}/\varepsilon_{\text{score},\infty})^2$. It yields

$$\mathsf{TV}(p_T, q) \lesssim \sqrt{\mathsf{KL}(q \,\|\, \gamma^d)}\exp(-T) + (\varepsilon_{\text{score},\infty} + L\sqrt{dh} + L\mathfrak{m}_2 h)\sqrt{T} + \frac{N\varepsilon^2_{\text{score}}}{\varepsilon^2_{\text{score},\infty}} .$$

Theorem 1 follows after optimizing over the choice of $\varepsilon_{\text{score},\infty}$. □

D.4 Proof of Corollary 2

Corollary 2 follows from Theorem 1 and the following regularization lemma.

Lemma 7. *Suppose that* $\operatorname{supp} q \subseteq \mathsf{B}(0, R)$ *where* $R \geq 1$*, 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. $W_2(q_t, q) \leq \varepsilon$.
2. q_t *satisfies*
$$\mathsf{KL}(q_t \,\|\, \gamma^d) \lesssim \frac{\sqrt{d}\,(R \vee \sqrt{d})^3}{\varepsilon^2}\,.$$
3. *For every* $t' \geq t$*,* $q_{t'}$ *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 $t \lesssim 1$,
$$\begin{aligned} W_2^2(q, q_t) &\leq \mathbb{E}\big[\big\|\big(1-\exp(-t)\big)\,\bar{X}_0 + \sqrt{1-\exp(-2t)}\,Z\big\|^2\big] \\ &= \big(1-\exp(-t)\big)^2\,\mathbb{E}[\|\bar{X}_0\|^2] + \big(1-\exp(-2t)\big)\,d \lesssim R^2t^2 + dt\,. \end{aligned}$$
We now take $t \lesssim \min\{\varepsilon/R, \varepsilon^2/d\}$ to ensure that $W_2^2(q, q_t) \leq \varepsilon^2$. 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
$$\mathsf{KL}(q_t \,\|\, \gamma^d) \leq \frac{W_2^2(q, \gamma^d)}{4t} \lesssim \frac{W_2^2(q, \delta_0) + W_2^2(\gamma^d, \delta_0)}{t} \lesssim \frac{\sqrt{d}\,(R \vee \sqrt{d})^3}{\varepsilon^2}\,.$$
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
$$\|\nabla^2 \ln q_{t'}\|_{\mathrm{op}} \lesssim \frac{1}{1-\exp(-2t')} \vee \frac{\exp(-2t')\,R^2}{(1-\exp(-2t'))^2} \lesssim \frac{1}{t} \vee \frac{R^2}{t^2} \lesssim \frac{dR^2\,(R \vee \sqrt{d})^2}{\varepsilon^4}\,.$$

□

D.5 Auxiliary lemmas

In this section, we prove some auxiliary lemmas which are used in the proof of Theorem 1.

Lemma 8 (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* $t \geq 0$*,*
$$\mathbb{E}[\|\bar{X}_t\|^2] \leq d \vee \mathfrak{m}_2^2\,.$$
2. *(score function bound) For all* $t \geq 0$*,*
$$\mathbb{E}[\|\nabla \ln q_t(\bar{X}_t)\|^2] \leq Ld\,.$$

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 9 (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 8. □

E Critically damped Langevin diffusion (CLD)

E.1 Background on the CLD

The critically damped Langevin diffusion (CLD) is based on the forward process
$$\begin{aligned}\mathrm{d}\bar{X}_t &= -\bar{V}_t\,\mathrm{d}t\,,\\ \mathrm{d}\bar{V}_t &= -(\bar{X}_t + 2\,\bar{V}_t)\,\mathrm{d}t + 2\,\mathrm{d}B_t\,.\end{aligned} \tag{E.1}$$

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 (E.1) 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 (B.1), the corresponding reverse process is
$$\begin{aligned}\mathrm{d}\bar{X}_t^{\leftarrow} &= -\bar{V}_t^{\leftarrow}\,\mathrm{d}t\,,\\ \mathrm{d}\bar{V}_t^{\leftarrow} &= \big(\bar{X}_t^{\leftarrow} + 2\,\bar{V}_t^{\leftarrow} + 4\,\nabla_v \ln \boldsymbol{q}_{T-t}(\bar{X}_t^{\leftarrow}, \bar{V}_t^{\leftarrow})\big)\,\mathrm{d}t + 2\,\mathrm{d}B_t\,,\end{aligned} \tag{E.2}$$

where $\boldsymbol{q}_t := \mathrm{law}(\bar{X}_t, \bar{V}_t)$ 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 $t \in [kh, (k+1)h]$.

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 $\bar{\boldsymbol{\theta}}_t \coloneqq (\bar{X}_t, \bar{V}_t)$, then the forward process satisfies the linear SDE

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

The solution to the SDE is given by

$$\bar{\boldsymbol{\theta}}_t = \exp(t\boldsymbol{A}_\gamma) \, \bar{\boldsymbol{\theta}}_0 + \int_0^t \exp\{(t-s) \, \boldsymbol{A}_\gamma\} \, \Sigma_\gamma \, \mathrm{d}B_s \,, \tag{E.3}$$

which means that by the Itô isometry,

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

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

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

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.

E.2 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 $t \geq 0$, the score $\nabla_v \ln \boldsymbol{q}_t$ is L-Lipschitz.*

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

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

If we ignore the dependence on L and assume that the score estimate is sufficiently accurate, then the iteration complexity guarantee of Theorem 1 is $N = \widetilde{\Theta}(d/\varepsilon^2)$. On the other hand, recall from Section E.1 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 10 (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 E.1) at time T, and suppose that the step size $h \coloneqq T/N$ satisfies $h \lesssim 1/L$, where $L \geq 1$. Then, there is a universal constant $c > 0$ such that*

$$\begin{aligned} &\mathsf{TV}(\boldsymbol{p}_T, q \otimes \gamma^d) \\ &\lesssim \underbrace{\sqrt{\mathsf{KL}(q \parallel \gamma^d) + \mathsf{FI}(q \parallel \gamma^d)} \exp(-cT)}_{\textit{convergence of forward process}} + \underbrace{(L\sqrt{dh} + L\mathfrak{m}_2 h)\sqrt{T}}_{\textit{discretization error}} + \underbrace{\frac{N^{1/3} \varepsilon_{\text{score}}^{2/3}}{T^{1/6}}}_{\textit{score estimation error}} \,, \end{aligned}$$

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

Proof. See Section F. □

Note that the result of Theorem 10 is in fact no better than our guarantee for DDPM in Theorem 1. 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 F.1, from the form of the reverse process (E.2), 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 1 and 10, 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 11. *Let $\boldsymbol{p}_T$ be the output of the SGM algorithm based on the CLD (Section E.1) 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 (E.2) respectively (see Section F for details), it holds that*

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

Proof. See Section F.4. □

Theorem 11 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 11 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 the DDPM algorithm.

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 $\nabla_v \ln \boldsymbol{q}_t$ for the SGM with the CLD is easier to estimate than the score function for DDPM, providing a *statistical* benefit of using the CLD. We do not investigate this question here and leave it as future work.

F Proofs for CLD

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.

F.1 Girsanov discretization argument

In order to apply Girsanov's theorem, we introduce the path measures $\boldsymbol{P}_T^{\infty,\boldsymbol{q}_T}$ 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

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

and

$$\sigma_t = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}.$$

Here, σ_t is not invertible, so strictly speaking Theorem 4 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 12. *Suppose that Novikov's condition (D.1) holds. Then,*

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

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

Theorem 13 (discretization error for CLD)**.** *Suppose that Assumptions 2, 4, and 5 hold. Let $\boldsymbol{Q}_T^{\leftarrow}$ and $\boldsymbol{P}_T^{\infty,\boldsymbol{q}_T}$ denote the measures on path space corresponding to the reverse process (E.2) and the SGM algorithm with L^∞-accurate score estimate initialized at $\boldsymbol{q}_T$. Assume that $L \geq 1$ and $h \lesssim 1/L$. Then, it holds that*

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

Similarly to Section D.2, we prove the theorem assuming that Novikov's condition (D.1) holds. This assumption can be lifted using similar arguments to the ones in Appendix G, and we omit the details.

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

$$\begin{aligned}&\|\boldsymbol{s}_{T-kh}^{\infty}(X_{kh}, V_{kh}) - \nabla_v \ln \boldsymbol{q}_{T-t}(X_t, V_t)\|^2 \\ &\quad\lesssim \|\boldsymbol{s}_{T-kh}^{\infty}(X_{kh}, V_{kh}) - \nabla_v \ln \boldsymbol{q}_{T-kh}(X_{kh}, V_{kh})\|^2 \\ &\qquad + \|\nabla_v \ln \boldsymbol{q}_{T-kh}(X_{kh}, V_{kh}) - \nabla_v \ln \boldsymbol{q}_{T-t}(X_{kh}, V_{kh})\|^2 \\ &\qquad + \|\nabla_v \ln \boldsymbol{q}_{T-t}(X_{kh}, V_{kh}) - \nabla_v \ln \boldsymbol{q}_{T-t}(X_t, V_t)\|^2 \\ &\quad\lesssim \varepsilon_{\mathrm{score},\infty}^2 + \left\|\nabla_v \ln \frac{\boldsymbol{q}_{T-kh}}{\boldsymbol{q}_{T-t}}(X_{kh}, V_{kh})\right\|^2 + L^2 \|(X_{kh}, V_{kh}) - (X_t, V_t)\|^2\,. \qquad \text{(F.1)}\end{aligned}$$

The change in the score function is bounded by Lemma 14, which generalizes [LLT22, Lemma C.12]. From the representation (E.3) 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 14, we deduce that if $h \lesssim 1/L$, then

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

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

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 (F.1). Hence,

$$\begin{aligned}&\|\boldsymbol{s}_{T-kh}^{\infty}(X_{kh}, V_{kh}) - \nabla_v \ln \boldsymbol{q}_{T-t}(X_t, V_t)\|^2 \\ &\quad\lesssim \varepsilon_{\mathrm{score},\infty}^2 + L^2 dh + L^2 h^2 \left(\|X_{kh}\|^2 + \|V_{kh}\|^2\right) \\ &\qquad + L^2 h^2 \|\nabla \ln \boldsymbol{q}_{T-t}(X_t, V_t)\|^2 + L^2 \|(X_{kh}, V_{kh}) - (X_t, V_t)\|^2\,.\end{aligned}$$

By applying the moment bounds in Lemma 15 together with Lemma 16 on the movement of the CLD, we obtain

$$\begin{aligned}
&\|\boldsymbol{s}^{\infty}_{T-kh}(X_{kh}, V_{kh}) - \nabla_v \ln \boldsymbol{q}_{T-t}(X_t, V_t)\|^2 \\
&\quad \lesssim \varepsilon^2_{\mathrm{score},\infty} + L^2 dh + L^2 h^2 \left(d + \mathfrak{m}_2^2\right) + L^3 dh^2 + L^2 \left(dh + \mathfrak{m}_2^2 h^2\right) \\
&\quad \lesssim \varepsilon^2_{\mathrm{score},\infty} + L^2 dh + L^2 \mathfrak{m}_2^2 h^2 \,.
\end{aligned}$$

Together with Corollary 12 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 6). 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

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

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

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

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 F.4, we use this observation in order to prove a rigorous lower bound against discretization of SGMs based on the CLD.

F.2 Proof of Theorem 10

Proof. [Proof of Theorem 10] Similarly to the proof of Theorem 1 in Section D.3, Theorem 13 together with the data processing inequality yield

$$\begin{aligned}
&\mathsf{TV}\big(\mathrm{law}(X_{kh}^{\leftarrow,\infty}, V_{kh}^{\leftarrow,\infty} : k = 0, 1, \ldots, N-1),\ \mathrm{law}(\bar{X}_{kh}^{\leftarrow}, \bar{V}_{kh}^{\leftarrow} : k = 0, 1, \ldots, N-1)\big) \\
&\quad \lesssim \sqrt{\mathsf{KL}(\boldsymbol{q}_T \parallel \boldsymbol{\gamma}^{2d})} + \left(\varepsilon_{\mathrm{score},\infty} + L\sqrt{dh} + L\mathfrak{m}_2 h\right)\sqrt{T} \,.
\end{aligned}$$

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,

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

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

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

Since $\boldsymbol{q}_0 = q \otimes \gamma^d$ and $\boldsymbol{\gamma}^{2d} = \gamma^d \otimes \gamma^d$, then $\mathcal{L}(\boldsymbol{q}_0 \parallel \boldsymbol{\gamma}^{2d}) \lesssim \mathsf{KL}(q \parallel \gamma^d) + \mathsf{FI}(q \parallel \gamma^d)$. Applying Lemma 3 and Pinsker's inequality, we deduce that

$$\begin{aligned}&\mathsf{TV}(\boldsymbol{p}_T, \boldsymbol{q}_0)\\ &\quad\lesssim \sqrt{\mathsf{KL}(q \parallel \gamma^d) + \mathsf{FI}(q \parallel \gamma^d)}\exp(-cT) + (\varepsilon_{\mathrm{score},\infty} + L\sqrt{dh} + L\mathfrak{m}_2 h)\sqrt{T} + \frac{N\varepsilon_{\mathrm{score}}^2}{\varepsilon_{\mathrm{score},\infty}^2}\,.\end{aligned}$$

Theorem 10 follows from optimizing over the choice of $\varepsilon_{\mathrm{score},\infty}$. □

F.3 Auxiliary lemmas

We begin with the perturbation lemma for the score function.

Lemma 14 (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 $\|\boldsymbol{M}_0 - \boldsymbol{I}_{2d}\|_{\mathrm{op}} \le \zeta$, 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 $L \le \frac{1}{4\|\boldsymbol{M}_1\|_{\mathrm{op}}}$. Then, it holds that*

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

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

$$\left\|\nabla \ln \frac{\boldsymbol{q} * \mathsf{normal}(0, \boldsymbol{M}_1)}{\boldsymbol{q}}(\boldsymbol{\theta})\right\| \lesssim L\sqrt{\|\boldsymbol{M}_1\|_{\mathrm{op}}\, d} + L\,\|\boldsymbol{M}_1\|_{\mathrm{op}}\,\|\nabla \boldsymbol{H}(\boldsymbol{\theta})\|\,. \tag{F.2}$$

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

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

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

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

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

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

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, [DKR19, 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 (F.2).

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

$$\begin{aligned}&\left\|\nabla \ln \frac{(\boldsymbol{M}_0)_\# \boldsymbol{q} * \mathsf{normal}(0, \boldsymbol{M}_1)}{\boldsymbol{q}}(\boldsymbol{\theta})\right\| \\ &\qquad \le \left\|\nabla \ln \frac{(\boldsymbol{M}_0)_\# \boldsymbol{q} * \mathsf{normal}(0, \boldsymbol{M}_1)}{(\boldsymbol{M}_0)_\# \boldsymbol{q}}(\boldsymbol{\theta})\right\| + \left\|\nabla \ln \frac{(\boldsymbol{M}_0)_\# \boldsymbol{q}}{\boldsymbol{q}}(\boldsymbol{\theta})\right\|.\end{aligned}$$

We can apply (F.2) 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

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

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

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

Combining the bounds,

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

so the lemma follows. □

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

Lemma 15 (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* (E.1)*.*

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
$$\mathbb{E}[\|(\bar{X}_t, \bar{V}_t)\|^2] = W_2^2(\boldsymbol{q}_t, \delta_{\boldsymbol{0}}) \lesssim W_2^2(\boldsymbol{q}_t, \boldsymbol{\gamma}^{2d}) + W_2^2(\boldsymbol{\gamma}^{2d}, \delta_{\boldsymbol{0}}) \lesssim d + W_2^2(\boldsymbol{q}_t, \boldsymbol{\gamma}^{2d})\,.$$
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]:
$$W_2^2(\boldsymbol{q}_t, \boldsymbol{\gamma}^{2d}) \lesssim W_2^2(\boldsymbol{q}, \boldsymbol{\gamma}^{2d}) \lesssim W_2^2(\boldsymbol{q}, \delta_{\boldsymbol{0}}) + W_2^2(\delta_{\boldsymbol{0}}, \boldsymbol{\gamma}^{2d}) \lesssim d + \mathfrak{m}_2^2\,.$$

2. The proof is the same as in Lemma 8.

□

Lemma 16 (movement bound for CLD)**.** *Suppose that Assumptions 2 holds. Let* $(\bar{X}_t, \bar{V}_t)_{t\in[0,T]}$ *denote the forward process* (E.1)*. For* $0 < s < t$ *with* $\delta \coloneqq t - s$*, 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 15. 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 15 again. □

F.4 Lower bound against CLD

Proof. [Proof of Theorem 11] Since $\boldsymbol{q}_0 = \gamma^d \otimes \gamma^d = \boldsymbol{\gamma}^{2d}$ is stationary for the forward process (E.1), 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 $\boldsymbol{P}_T = \boldsymbol{P}_T^{\infty,\boldsymbol{q}_T}$. From Girsanov's theorem in the form of Corollary 12 and from $\boldsymbol{s}_{T-kh}(x,v) = \nabla_v \ln \boldsymbol{q}_{T-kh}(x,v) = -v$, we have

$$\mathsf{KL}(\boldsymbol{Q}_T^{\leftarrow} \parallel \boldsymbol{P}_T) = 2\sum_{k=0}^{N-1} \mathbb{E}_{\boldsymbol{Q}_T^{\leftarrow}} \int_{kh}^{(k+1)h} \|V_{kh} - V_t\|^2\,\mathrm{d}t\,. \tag{F.3}$$

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 (F.3),

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

This proves the result. □

This lower bound shows that the Girsanov discretization argument of Theorem 13 is essentially tight (except possibly the dependence on L).

G Novikov's condition for DDPM

In Section D.2, we proved the Girsanov discretization bound for DDPM (Theorem 6) under the assumption that the path measures P_T^{∞,q_T} and $Q_T^{\leftarrow}$ satisfy Novikov's condition (D.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 $Q_T^{\leftarrow}$.

G.1 Truncation argument

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 $\phi_R(x) \coloneqq \psi(\|x\|/R)$. 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. □

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

$$\mathrm{d}\bar{X}_t^{\mathsf{tr},\leftarrow} = \phi_R(\bar{X}_t^{\mathsf{tr},\leftarrow})\,\{\bar{X}_t^{\mathsf{tr},\leftarrow} + 2\,\nabla \ln q_{T-t}(\bar{X}_t^{\mathsf{tr},\leftarrow})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \tag{G.1}$$

$$\mathrm{d}X_t^{\mathsf{tr},\leftarrow,\infty,q_T} = \phi_R(X_t^{\mathsf{tr},\leftarrow,\infty,q_T})\,\{X_t^{\mathsf{tr},\leftarrow,\infty,q_T} + 2\,s_{T-kh}^{\infty}(X_{kh}^{\mathsf{tr},\leftarrow,\infty,q_T})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, \tag{G.2}$$

where $\bar{X}_0^{\mathsf{tr},\leftarrow} = X_0^{\mathsf{tr},\leftarrow,\infty,q_T}$ is obtained by sampling $\bar{X}_T \sim q_T$ and setting $\bar{X}_0^{\mathsf{tr},\leftarrow} = \bar{X}_T$ if $\|\bar{X}_T\| \le R$ and setting $\bar{X}_0^{\mathsf{tr},\leftarrow} = 0$ otherwise. For clarity of presentation, we suppress the dependence on R in the notation.

Let $q_t^{\mathsf{tr},\leftarrow} := \mathrm{law}(\bar{X}_t^{\mathsf{tr},\leftarrow})$ and $p_t^{\mathsf{tr},\infty,q_T} := \mathrm{law}(X_t^{\mathsf{tr},\leftarrow,\infty,q_T})$. Likewise, denote the corresponding measures over path space by $Q_T^{\mathsf{tr},\leftarrow}$ and $P_T^{\mathsf{tr},\infty,q_T}$. As it is convenient to think about a single sochastic process, for ease of notation we denote this by $(X_t^{\mathsf{tr}})_{t\in[0,T]}$ so that under $Q_T^{\mathsf{tr},\leftarrow}$ (resp. $P_T^{\mathsf{tr},\infty,q_T}$), $(X_t^{\mathsf{tr}})_{t\in[0,T]}$ has the law of the modified reverse process (G.1) (resp. the modified SGM process (G.2)). We first verify that these modified processes satisfy Novikov's condition.

Lemma 19. *It holds that*

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

Proof. For any x,

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

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

We can thus apply Girsanov's theorem (Theorem 4), with the choices $P_T = P_T^{\mathsf{tr},\infty,q_T}$, $Q_T = Q_T^{\mathsf{tr},\leftarrow}$, $\tilde{b}_t = \phi_R(X_t^{\mathsf{tr}})\,\{X_t^{\mathsf{tr}} + 2\, s_{T-kh}^{\infty}(X_{kh}^{\mathsf{tr}})\}$, $b_t = \phi_R(X_t^{\mathsf{tr}})\,\{X_t^{\mathsf{tr}} + 2\,\nabla \ln q_{T-t}(X_t^{\mathsf{tr}})\}$, and $\sigma_t = \sqrt{2}\, I_d$. It yields the following corollary.

Corollary 20. *It holds that*

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

Our analysis for the quantity on the right-hand side of Corollary 20 proceeds along similar lines to the argument in Section D.2, though the moment and movement bounds require some additional work as we must work with $\bar{X}_t^{\mathsf{tr},\leftarrow}$ 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 $Q_T^{\mathsf{tr},\leftarrow}$ and $P_T^{\mathsf{tr},\infty,q_T}$ denote the measures on path space corresponding to the modified processes (G.1) and (G.2) respectively. Assume that $L \ge 1$ and $h \lesssim 1/L$. Then,*

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

Proof. First, as before, we can decompose

$$\begin{aligned}&\|\phi_R(X_{kh}^{\mathsf{tr}})\, s_{T-kh}^{\infty}(X_{kh}^{\mathsf{tr}}) - \phi_R(X_t^{\mathsf{tr}})\, \nabla \ln q_{T-t}(X_t^{\mathsf{tr}})\|^2 \\ &\quad\lesssim \|\phi_R(X_{kh}^{\mathsf{tr}})\,\{s_{T-kh}^{\infty}(X_{kh}^{\mathsf{tr}}) - \nabla \ln q_{T-kh}(X_{kh}^{\mathsf{tr}})\}\|^2 \\ &\qquad\quad + \|\phi_R(X_{kh}^{\mathsf{tr}})\,\{\nabla \ln q_{T-kh}(X_{kh}^{\mathsf{tr}}) - \nabla \ln q_{T-t}(X_{kh}^{\mathsf{tr}})\}\|^2 \\ &\qquad\quad + \|\phi_R(X_{kh}^{\mathsf{tr}})\, \nabla \ln q_{T-t}(X_{kh}^{\mathsf{tr}}) - \phi_R(X_t^{\mathsf{tr}})\, \nabla \ln q_{T-t}(X_t^{\mathsf{tr}})\|^2 \\ &\quad\lesssim \varepsilon_{\mathrm{score},\infty}^2 + \left\|\phi_R(X_{kh}^{\mathsf{tr}})\, \nabla \ln \frac{q_{T-kh}}{q_{T-t}}(X_{kh}^{\mathsf{tr}})\right\|^2 + L'^2\, \|X_{kh}^{\mathsf{tr}} - X_t^{\mathsf{tr}}\|^2\,,\end{aligned}$$

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

Because (D.3) and (D.4) hold pointwise, we can bound the second term above by

$$\begin{aligned}&\left\|\phi_R(X_{kh}^{\mathsf{tr}})\, \nabla \ln \frac{q_{T-kh}}{q_{T-t}}(X_{kh}^{\mathsf{tr}})\right\|^2 \\ &\quad\lesssim \phi_R(X_{kh}^{\mathsf{tr}})^2 \left(L^2 dh + L^2 h^2\, \|X_{kh}^{\mathsf{tr}}\|^2 + L^2 h^2\, \|\nabla \ln q_{T-t}(X_{kh}^{\mathsf{tr}})\|^2\right) \\ &\quad\lesssim L^2 dh + \phi_R(X_{kh}^{\mathsf{tr}})^2\, L^2 h^2\, \|X_{kh}^{\mathsf{tr}}\|^2 \\ &\qquad + \phi_R(X_t^{\mathsf{tr}})^2\, L^2 h^2\, \|\nabla \ln q_{T-t}(X_t^{\mathsf{tr}})\|^2 + L'^2\, \|X_{kh}^{\mathsf{tr}} - X_t^{\mathsf{tr}}\|^2\,,\end{aligned}$$

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

$$\begin{aligned}
&\|\phi_R(X^{\mathsf{tr}}_{kh})\, s^{\infty}_{T-kh}(X^{\mathsf{tr}}_{kh}) - \phi_R(X^{\mathsf{tr}}_t)\, \nabla \ln q_{T-t}(X^{\mathsf{tr}}_t)\|^2 \\
&\quad \lesssim \varepsilon^2_{\mathrm{score},\infty} + L^2 dh + \phi_R(X^{\mathsf{tr}}_{kh})^2\, L^2 h^2\, \|X^{\mathsf{tr}}_{kh}\|^2 \\
&\qquad + \phi_R(X^{\mathsf{tr}}_t)^2\, L^2 h^2\, \|\nabla \ln q_{T-t}(X^{\mathsf{tr}}_t)\|^2 + L'^2\, \|X^{\mathsf{tr}}_{kh} - X^{\mathsf{tr}}_t\|^2 \,.
\end{aligned}$$

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

$$\begin{aligned}
&\limsup_{R\to\infty} \mathbb{E}_{Q^{\mathsf{tr},\leftarrow}_T}[\|\phi_R(X^{\mathsf{tr}}_{kh})\, s^{\infty}_{T-kh}(X^{\mathsf{tr}}_{kh}) - \phi_R(X^{\mathsf{tr}}_t)\, \nabla \ln q_{T-t}(X^{\mathsf{tr}}_t)\|^2] \\
&\quad \lesssim \varepsilon^2_{\mathrm{score},\infty} + L^2 dh + L^2 h^2\, \mathbb{E}_{Q^{\leftarrow}_T}[\|X_t\|^2] \\
&\qquad + L^2 h^2\, \mathbb{E}_{Q^{\leftarrow}_T}[\|\nabla \ln q_{T-t}(X_t)\|^2] + L^2\, (\mathfrak{m}_2^2 h^2 + dh) \,.
\end{aligned}$$

We can bound $\mathbb{E}_{Q^{\leftarrow}_T}[\|X_t\|^2] \le d \vee \mathfrak{m}_2^2$ and $\mathbb{E}_{Q^{\leftarrow}_T}[\|\nabla \ln q_{T-t}(X_t)\|^2] \le Ld$ via the moment bounds of Lemma 8.

Putting everything together, we get

$$\limsup_{R\to\infty} \mathbb{E}[\|\phi_R(X^{\mathsf{tr}}_{kh})\, s^{\infty}_{T-kh}(X^{\mathsf{tr}}_{kh}) - \phi_R(X^{\mathsf{tr}}_t)\, \nabla \ln q_{T-t}(X^{\mathsf{tr}}_t)\|^2] \lesssim \varepsilon^2_{\mathrm{score},\infty} + L^2 dh + L^2 \mathfrak{m}_2^2 h^2 \,.$$

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 6] Consider a coupling of the laws of $(X_t)_{t\in[0,T]}$ and $(X^{\mathsf{tr}}_t)_{t\in[0,T]}$ under $Q^{\leftarrow}_T$ (resp. P^{∞,q_T}_T) and $Q^{\mathsf{tr},\leftarrow}_T$ (resp. $P^{\mathsf{tr},\infty,q_T}_T$) such that if $X_t \in \mathsf{B}(0,R)$ for all $t \in [0,T]$, then $X^{\mathsf{tr}}_t = X_t$ for $t \in [0,T]$. Denote this event by $\mathcal{E}^R$ (resp. $\mathcal{E}'^R$). Then

$$\mathsf{TV}(Q^{\mathsf{tr},\leftarrow}_T, Q^{\leftarrow}_T) \le \mathbb{P}[\mathcal{E}^R] \qquad \text{and} \qquad \mathsf{TV}(P^{\mathsf{tr},\infty,q_T}_T, P^{\infty,q_T}_T) \le \mathbb{P}[\mathcal{E}'^R] \,.$$

Clearly $\lim_{R\to\infty} \mathbb{P}[\mathcal{E}^R] = 0$ and $\lim_{R\to\infty} \mathbb{P}[\mathcal{E}'^R] = 0$. By Theorem 21 and Pinsker's inequality,

$$\limsup_{R\to\infty} \mathsf{TV}(Q^{\mathsf{tr},\leftarrow}_T, P^{\mathsf{tr},\infty,q_T}_T)^2 \lesssim (\varepsilon^2_{\mathrm{score},\infty} + L^2 dh + L^2 \mathfrak{m}_2^2 h^2)\, T \,.$$

Hence, by the triangle inequality,

$$\begin{aligned}
\mathsf{TV}(Q^{\leftarrow}_T, P^{\infty,q_T}_T) &\le \limsup_{R\to\infty} \{\mathsf{TV}(Q^{\mathsf{tr},\leftarrow}_T, Q^{\leftarrow}_T) + \mathsf{TV}(P^{\mathsf{tr},\infty,q_T}_T, P^{\infty,q_T}_T) + \mathsf{TV}(Q^{\mathsf{tr},\leftarrow}_T, P^{\mathsf{tr},\infty,q_T}_T)\} \\
&\lesssim (\varepsilon_{\mathrm{score},\infty} + L\sqrt{dh} + L\mathfrak{m}_2 h)\, \sqrt{T} \,.
\end{aligned}$$

This is the claimed result. □

G.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,*

$$\mathbb{P}\{\bar{X}^{\leftarrow}_t \in \mathsf{B}(0,R) \textit{ for all } t \in [0,T]\} \ge 1 - o\big(\frac{1}{R^2}\big) \qquad \textit{as } R \to \infty \,.$$

Proof. As $\bar{X}^{\leftarrow}$ is the reverse of the OU process, we can equivalently bound

$$\mathbb{P}\{\bar{X}_t \in \mathsf{B}(0,R) \text{ for all } t \in [0,T]\} \,.$$

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 $\lambda = 1/(4\tilde{T})$,

$$\mathbb{P}\{\sup_{t\in[0,\tilde{T}]} \|\tilde{B}_t\| > \frac{R}{2\sqrt{2}}\} \le 3^d \exp\big(-\frac{R^2}{32\tilde{T}}\big) = o\big(\frac{1}{R^2}\big)\,.$$

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

Lemma 23 (effect of truncation on the moment bound)**.** *Under Assumption 2, we have*

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

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

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

Because $\phi_R(X_t^{\mathsf{tr}}) = 0$ if $\|X_t^{\mathsf{tr}}\| \ge 2R$, 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 24 (effect of truncation on the score function bound)**.** *Under Assumptions 1 and 2,*

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

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

Proof. Consider the same coupling of the laws of $(X_t)_{t\in[0,T]}$ and $(X_t^{\mathsf{tr}})_{t\in[0,T]}$ from the proof of Lemma 23, 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

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

Because $\phi_R(X_t^{\mathsf{tr}}) = 0$ if $\|X_t^{\mathsf{tr}}\| \ge 2R$, 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 25 (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*

$$\limsup_{R\to\infty} \mathbb{E}_{Q_T^{\mathsf{tr},\leftarrow}}[\|\bar{X}_t^{\mathsf{tr},\leftarrow} - \bar{X}_s^{\mathsf{tr},\leftarrow}\|^2] \lesssim \mathfrak{m}_2^2 h^2 + dh\,.$$

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

$$\begin{aligned}
&\mathbb{E}[\|\bar{X}_t^{\mathsf{tr},\leftarrow} - \bar{X}_s^{\mathsf{tr},\leftarrow}\|^2] \\
&\quad= \mathbb{E}\Big[\Big\|\int_s^t \phi_R(\bar{X}_r^{\mathsf{tr},\leftarrow})\,\{\bar{X}_r^{\mathsf{tr},\leftarrow} + 2\,\nabla \ln q_{T-r}(\bar{X}_r^{\mathsf{tr},\leftarrow})\}\,\mathrm{d}r + \sqrt{2}\,(B_t - B_s)\Big\|^2\Big] \\
&\quad\lesssim h \int_s^t \mathbb{E}[\phi_R(\bar{X}_r^{\mathsf{tr},\leftarrow})^2\, \|\bar{X}_r^{\mathsf{tr},\leftarrow}\|^2]\,\mathrm{d}r \\
&\qquad\quad + h \int_s^t \mathbb{E}[\phi_R(\bar{X}_r^{\mathsf{tr},\leftarrow})^2\, \|\nabla \ln q_{T-r}(\bar{X}_r^{\mathsf{tr},\leftarrow})\|^2]\,\mathrm{d}r + dh\,.
\end{aligned}$$

Taking the lim sup on both sides as $R \to \infty$, we can apply Lemmas 23 and 24 to get

$$\limsup_{R\to\infty} \mathbb{E}[\|\bar{X}_t^{\mathsf{tr},\leftarrow} - \bar{X}_s^{\mathsf{tr},\leftarrow}\|^2] \lesssim h \int_s^t \mathbb{E}[\|\bar{X}_r^{\leftarrow}\|^2]\,\mathrm{d}r + h \int_s^t \mathbb{E}[\|\nabla \ln q_{T-r}(\bar{X}_r^{\leftarrow})\|^2]\,\mathrm{d}r + dh\,.$$

The claimed bound then follows from the moment and score function bounds of Lemma 8. □