Convergence of score-based generative modeling for general data distributions

Holden Lee
Department of Applied Mathematics and Statistics
Johns Hopkins University
hlee283@jhu.edu

Jianfeng Lu
Department of Mathematics
Duke University
jianfeng@math.duke.edu

Yixin Tan
Department of Mathematics
Duke University
yixin.tan@duke.edu

Abstract

We give polynomial convergence guarantees for denoising diffusion models for general data distributions, with no assumptions related to functional inequalities or smoothness. Assuming L^2-accurate score estimates, we obtain Wasserstein distance guarantees for *any* distribution of bounded support or sufficiently decaying tails, as well as TV guarantees for distributions with further smoothness assumptions.

1 Introduction

Diffusion models have gained huge popularity in recent years in machine learning, as a method to learn and generate new samples from a data distribution. Score-based generative modeling (SGM), as a particular kind of diffusion model, uses learned score functions (gradients of the log-pdf) to transform white noise to the data distribution through following a stochatic differential equation. While SGM has achieved state-of-the-art performance for artificial image and audio generation [SE19; Dat+19; Gra+19; SE20; Son+20; Men+21; Son+21b; Son+21a; Jin+22], including being a key component of text-to-image systems [Ram+22], our theoretical understanding of these models is still nascent.

In particular, basic questions on the convergence of the generated distribution to the data distribution remain unanswered. Recent theoretical work on SGM has attempted to answer these questions [De +21; LLT22; De 22], but they either suffer from exponential dependence on parameters or rely on strong assumptions on the data distribution such as functional inequalities or smoothness, which are rarely satisfied in practical situations. Indeed, data distributions—such as those arising in the hallmark application of generating images from text—are often multimodal and hence do not satisfy functional inequalities with reasonable constants, and supported on lower-dimensional manifolds and hence not smooth.

In this work, we aim to develop theoretical convergence guarantees with polynomial complexity for SGM under minimal data assumptions. We refer to Appendix A for a thorough discussion about prior works and our contributions.

NeurIPS 2022 Workshop on Score-Based Methods.

2 Problem setting

Given samples from a data distribution P_{data}, the problem of generative modeling is to learn the distribution in a way that allows generation of new samples. A general framework for many score-based generative models is where noise is injected into P_{data} via a forward SDE [Son+20]

$$d\widetilde{x}_t = f(\widetilde{x}_t, t)\, dt + g(t)\, dw_t, \quad t \in [0, T], \tag{1}$$

where $\widetilde{x}_0 \sim \widetilde{P}_0 := P_{\text{data}}$. Let $\widetilde{p}_t$ denote the density of $\widetilde{x}_t$. Remarkably, $\widetilde{x}_t$ also satisfies a reverse-time SDE,

$$d\widetilde{x}_t = [f(\widetilde{x}_t, t) - g(t)^2 \nabla \ln \widetilde{p}_t(\widetilde{x}_t)]\, dt + g(t)\, d\widetilde{w}_t, \quad t \in [0, T], \tag{2}$$

where $\widetilde{w}_t$ is a backward Brownian motion [And82]. Because the forward process transforms the data distribution to noise, the hope is to use the backwards process to transform noise into samples.

Of course, in practice the score function $\nabla \ln \widetilde{p}_t$ is not available, and and has to be learned from empirical samples $\widetilde{x}_t$ from the forward SDE starting from the data samples. Denoising auto-encoders are often used for the score function estimate, which will result in a guarantee that the L^2 error $\mathbb{E}_{\widetilde{p}_t}[\|\nabla \ln \widetilde{p}_t(x) - s(x,t)\|^2]$ is small [Vin11]. The natural question is then the following:

> Given L^2-error bounds of the score function, how close is the distribution generated by (2) (with score estimate $s(x,t)$ in place of $\nabla \ln \widetilde{p}_t$, and appropriate discretization) to the data distribution P_{data}?

We note it is more realistic to consider L^2 rather than L^∞-error, and this makes the analysis more challenging. In addition, we hope to obtain a result with as few structural assumptions as possible on P_{data}, so that it can be useful in realistic scenarios where SGM is applied.

3 Main results

To state our results, we will consider a specific type of SGM called denoising diffusion probabilistic modeling (DDPM) [HJA20], where in the forward SDE (1), $f(x,t) = -\frac{1}{2} g(t)^2 x$ for some non-decreasing function g to be chosen. The forward process is an Ornstein-Uhlenbeck process with time rescaling: $\widetilde{x}_t$ has the same distribution as

$$m_t \widetilde{x}_0 + \sigma_t z, \text{ where}$$
$$m_t = \exp\left[-\frac{1}{2}\int_0^t g(s)^2\, ds\right],\ \sigma_t^2 = 1 - \exp\left[-\int_0^t g(s)^2\, ds\right], \text{ and } z \sim N(0, I). \tag{3}$$

Given an estimate score function $s(x,t)$ approximating $\nabla \ln \widetilde{p}_t(x)$, we can simulate the reverse process (reparameterizing $t \hookleftarrow T - t$ and denoting $p_t := \widetilde{p}_{T-t}$)

$$dx_t = \frac{1}{2} g(T-t)^2 \left(x_t + 2\nabla \ln p_t(x_t)\right)\, dt + g(T-t)\, dw_t \tag{4}$$

with the exponential integrator discretization [ZC22], where $h_k = t_{k+1} - t_k$ and $\eta_{k+1} \sim N(0, I_d)$:

$$z_{t_{k+1}} = z_{t_k} + \gamma_{1,k}(z_{t_k} + 2s(T - t_k, z_{t_k})) + \sqrt{\gamma_{2,k}} \cdot \eta_{k+1}, \tag{5}$$

$$\text{where } \gamma_{1,k} = \exp\left[\frac{1}{2} G_{t_k, t_{k+1}}\right] - 1,\ \gamma_{2,k} = \exp\left[G_{t_k, t_{k+1}}\right] - 1, \text{ and } G_{t',t} := \int_{t'}^t g(T-s)^2\, ds. \tag{6}$$

We initialize z_0 with a prior distribution that approximates $p_0 = \widetilde{p}_T$ for sufficiently large T:

$$z_0 \sim q_0 = p_{\text{prior}} := N(0, \sigma_T^2 I_d) \approx N(0, I_d). \tag{7}$$

While we focus on DDPM, we note that the continuous process underlying DDPM is equivalent to that of score-matching Langevin diffusion (SMLD) under reparameterization in time and space (see [LLT22, §C.2]). We will further take $g \equiv 1$ for convenience in stating our results.

Our goal is to obtain a quantitative guarantee for the distance between the distribution q_{t_K} for z_{t_K} (for appropriate $t_K \approx T$) and P_{data}, under a L^2-score error guarantee. In the following, we assume a sequence of discretization points $0 = t_0 < t_1 < \cdots < t_K \le T$ has been chosen.

Assumption 1 (L^2 score error)**.** *For any $t \in \{T - t_0, \dots, T - t_K\}$, the error in the score estimate is bounded in $L^2(\widetilde{p}_t)$:*

$$\|\nabla \ln \widetilde{p}_t - s(\cdot, t)\|^2_{L^2(\widetilde{p}_t)} = \mathbb{E}_{\widetilde{p}_t}[\|\nabla \ln \widetilde{p}_t(x) - s(x,t)\|^2] \le \varepsilon_t^2 := \frac{\varepsilon_\sigma^2}{\sigma_t^4}.$$

We note that the gradient $\nabla \ln \widetilde{p}_t$ grows as $\frac{1}{\sigma_t^2}$ as $t \to 0$, so this is a reasonable assumption, and quantitatively weaker than a uniform bound over t.

Assumption 2 (Bounded support)**.** *P_{data} is supported on $B_R(0) := \{x \in \mathbb{R}^d : \|x\| \le R\}$.*

For simplicity, we assume bounded support when stating our main theorems, but note that our results generalize to distributions with sufficiently fast power decay. In the application of image generation, pixel values are bounded, so Assumption 2 is satisfied with R typically on the order of $\sqrt{d}$.

These are the only assumptions we need to obtain a polynomial complexity guarantee. We also consider the following stronger smoothness assumption, which is Assumption A.6 in [De 22] and will give better dependencies. Note that [De 22, Theorem I.8] shows a (nonuniform) version of Assumption 3 holds when p_0 is a smooth density on a convex submanifold.

Assumption 3. *The following bound of the Hessian of the log-pdf holds for any $t > 0$ and x:*

$$\left\|\nabla^2 \ln p_t(x)\right\| \le \frac{C}{\sigma_t^2},$$

for some constant $C > 0$.

Finally, the following smoothness assumption on $\widetilde{p}_0$ will allow us to obtain TV guarantees.

Assumption 4. *P_{data} admits a density $\widetilde{p}_0 \propto e^{-V(x)}$ where $V(x)$ is L-smooth.*

We are now ready to state our main theorems.

Algorithm 1 DDPM with exponential integrator [Son+20; ZC22]

INPUT: Time T; discretization points $0 = t_0 < t_1 < \cdots < t_N \le T$; score estimates $s(\cdot, T - t_k)$; radius R; function g (default: $g \equiv 1$)
Draw $z_0 \sim p_{\text{prior}}$ from the prior distribution p_{prior} given by (7).
for k from 1 to N **do**
  Compute z_{t_k} from $z_{t_{k-1}}$ using (5).
end for
Let $\widehat{z}_{t_N} = m_{T-t_N}^{-1} z_{t_N}$ if $z_{t_N} \in B_R(0)$; otherwise, let $\widehat{z}_{t_N} = 0$.

Theorem 3.1 (Wasserstein+TV error for distributions with bounded support)**.** *Suppose that Assumption 1 and 2 hold with $R \ge \sqrt{d}$. Then there is a sequence of discretization points $0 = t_0 < t_1 < \cdots < t_N < T$ with $N = O(\text{poly}(d, R, 1/\varepsilon_{\text{TV}}, 1/\varepsilon_{\text{W}}))$ such that if $\varepsilon_\sigma = \widetilde{O}\left(\frac{\varepsilon_{\text{TV}}^{6.5}\varepsilon_{\text{W}}^5}{R^9 d^{4.75}}\right)$, then the distribution of the scaled output $m_{T-t_N}^{-1} z_{t_N}$ of DDPM is ε_{TV}-close in TV distance to a distribution that is ε_W in W_2-distance from P_{data}. If in addition Assumption 3 holds with $C \ge R^2$, it suffices for $\varepsilon_\sigma = \widetilde{O}\left(\frac{\varepsilon_{\text{TV}}^4}{C^2 d}\right)$ (note that the $\widetilde{O}(\cdot)$ hides logarithmic dependence on ε_{W}).*

This result is perhaps surprising at first glance, as it is well known that for sampling algorithms such as Langevin Monte Carlo, structural assumptions on the target distribution—such as a log-Sobolev inequality—are required to obtain similar theoretical guarantees, even with the knowledge of the exact score function. The key reason that we can do better is that we utilize a *sequence* of score functions s_t along the reverse SDE, which is not available in standard sampling settings. Moreover, we choose T large enough so that $q_0 = p_{\text{prior}}$ is close to p_0, and it suffices to track the evolution of the true process (2), that is, maintain rather than decrease the error. To some extent, this result shows the power of DDPM and other reverse SDE-based methods compared with generative modeling based on standard Langevin Monte Carlo.

A statement with more precise dependencies, and which works for unbounded distributions with sufficiently decaying tails, can be found as Theorem E.2. We note that under the Hessian bound

(Assumption 3), up to logarithmic factors, the same score error bound suffices to obtain a fixed TV distance to a distribution arbitrarily close in W_2 distance. By truncating the resulting distribution, we can also obtain purely Wasserstein error bounds.

Theorem 3.2 (Wasserstein error for distributions with bounded support)**.** *In the same setting as Theorem 3.1, consider the distribution $\widehat{q}_{t_N}$ of the scaled and truncated output $\widehat{x}_{t_N}$ of DDPM. If Assumptions 1 and 2 hold with $R \geq \sqrt{d}$ and $\varepsilon_\sigma = \widetilde{O}\left(\frac{\varepsilon_{\mathrm{W}}^{18}}{R^{22}d^{4.75}}\right)$, then with appropriate (polynomial) choice of parameters, $W_2(\widehat{q}_{t_K}, P_{\text{data}}) \leq \varepsilon_{\mathrm{W}}$. If in addition Assumption 3 holds with $C \geq R^2$, then $\varepsilon_\sigma = \widetilde{O}\left(\frac{\varepsilon_{\mathrm{W}}^{8}}{C^2R^8d}\right)$ suffices.*

With an extra assumption on the smoothness of P_{data}, we can also obtain purely TV error bounds:

Theorem 3.3 (TV error for distributions under smoothness assumption)**.** *Suppose that Assumptions 1 and 4 hold, P_{data} is subexponential (with a fixed constant), and denote $R = \max\left\{\sqrt{d}, \mathbb{E}_{P_{data}}\|X\|\right\}$. Then there is a sequence of discretization points $0 = t_0 < t_1 < \cdots < t_N < T$ with $N = O(\mathrm{poly}(d, R, 1/\varepsilon_{\mathrm{TV}}))$ such that if $\varepsilon_\sigma = \widetilde{O}\left(\frac{\varepsilon_{\mathrm{TV}}^{11.5}}{R^{14}d^{2.25}L^5}\right)$, then the distribution q_{t_N} of the output z_{t_N} of DDPM satisfies $\mathrm{TV}(q_{t_N}, P_{\text{data}}) \leq \varepsilon_{\mathrm{TV}}$. If in addition Assumption 3 holds with $C \geq R^2$, then $\varepsilon_\sigma = \widetilde{O}\left(\frac{\varepsilon_{\mathrm{TV}}^{4}}{C^2d}\right)$ suffices.*

A more precise statement can be found as Theorem E.3, which also works more generally with sufficient tail decay. We note that this result can be derived directly by combining Theorem E.2 and a TV error bound between P_{data} and p_{t_N} (Lemma D.4) depending on the smoothness of P_{data}.

4 Proof overview

Our proof uses the framework by [LLT22] to convert guarantees under L^∞-accurate score function to under L^2-accurate score function. For the analysis under L^∞-accurate score function, we interpolate the discrete process with estimated score, $z_t \sim q_t$, and derive a differential inequality

$$\frac{d}{dt}\chi^2(q_t||p_t) = -g(T-t)^2\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 2\mathbb{E}\left[\left\langle g(T-t)^2(s(z_{t_k}, T-t_k) - \nabla \ln p_t(z_t), \nabla\frac{q_t(x)}{p_t(x)}\right\rangle\right].$$

We bound resulting error terms, making ample use of the Donsker-Varadhan variational principle to convert expectations to be under p_t. Under small enough step sizes, this shows that $\chi^2(q_t||p_t)$ grows slowly (Theorem B.10), which suffices as χ^2-divergence decays exponentially in the forward process.

The most challenging error term to deal with is the KL divergence term $\mathrm{KL}(q_t\psi_t||p_t)$. Our main innovation over the analysis of [LLT22] is bounding this term without a global log-Sobolev inequality for p_t. We note that it suffices for p_t to be a mixture of distributions each satisfying a log-Sobolev inequality, with the logarithm of the minimum mixture weight bounded below, and in Lemma C.2, we show that we can decompose any distribution of bounded support in this manner if we move a small amount of its mass.

In Section D, we show that this does not significantly affect the estimate of the score function, by interpreting the score function as solving a Bayesian inference problem: that of de-noising a noised data point. More precisely, we show in Lemma D.5 that the difference between the score functions of two different distributions can be bounded in L^2 in terms of their χ^2-divergence, which may be of independent interest.

Finally, we reduce from the L^2 to L^∞ setting by bounding the probabilities of hitting a bad set where the score error is large, and carefully choose parameters (Section E). This gives a TV error bound to $\widetilde{p}_\delta$—the forward distribution at small positive time. Finally, we can bound the Wasserstein distance of $\widetilde{p}_\delta$ to $\widetilde{P}_0$ (in the general case) or the TV distance (under additional smoothness of $\widetilde{P}_0$.)

In Section F we show that the Hessian is always bounded by $O\left(\frac{d}{\sigma_t^2}\right)$ with high probability (cf. Assumption 3). We speculate that a high-probability rather than uniform bound on the Hessian (as in Lemma B.13) can be used to obtain better dependencies, and leave this as an open problem.

References

[And82] Brian DO Anderson. "Reverse-time diffusion equation models". In: *Stochastic Processes and their Applications* 12.3 (1982), pp. 313–326.

[BMR20] Adam Block, Youssef Mroueh, and Alexander Rakhlin. "Generative modeling with denoising auto-encoders and Langevin sampling". In: *arXiv preprint arXiv:2002.00107* (2020).

[CCN21] Hong-Bin Chen, Sinho Chewi, and Jonathan Niles-Weed. "Dimension-free log-Sobolev inequalities for mixture distributions". In: *Journal of Functional Analysis* 281.11 (2021), p. 109236.

[Che+21] Sinho Chewi, Murat A Erdogdu, Mufan Bill Li, Ruoqi Shen, and Matthew Zhang. "Analysis of Langevin Monte Carlo from Poincaré to Log-Sobolev". In: *arXiv preprint arXiv:2112.12662* (2021).

[Che+22] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. *Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions*. arXiv:2209.11215. 2022.

[Dat+19] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. "Plug and play language models: A simple approach to controlled text generation". In: *arXiv preprint arXiv:1912.02164* (2019).

[De +21] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. "Diffusion Schrödinger bridge with applications to score-based generative modeling". In: *Advances in Neural Information Processing Systems* 34 (2021).

[De 22] Valentin De Bortoli. "Convergence of denoising diffusion models under the manifold hypothesis". In: *arXiv preprint arXiv:2208.05314* (2022).

[EHZ21] Murat A. Erdogdu, Rasa Hosseinzadeh, and Matthew S. Zhang. "Convergence of Langevin Monte Carlo in Chi-Squared and Renyi Divergence". In: *arXiv preprint arXiv:2007.11612* (2021).

[Gra+19] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. "Your classifier is secretly an energy based model and you should treat it like one". In: *arXiv preprint arXiv:1912.03263* (2019).

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851.

[Jin+22] Bowen Jing, Gabriele Corso, Renato Berlinghieri, and Tommi Jaakkola. "Subspace Diffusion Generative Models". In: *arXiv preprint arXiv:2205.01490* (2022).

[LLT22] Holden Lee, Jianfeng Lu, and Yixin Tan. "Convergence for score-based generative modeling with polynomial complexity". In: *arXiv preprint arXiv:2206.06227* (2022).

[LM00] Beatrice Laurent and Pascal Massart. "Adaptive estimation of a quadratic functional by model selection". In: *Annals of Statistics* (2000), pp. 1302–1338.

[Men+21] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. "SDEdit: Guided image synthesis and editing with stochastic differential equations". In: *International Conference on Learning Representations*. 2021.

[Ram+22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. "Hierarchical text-conditional image generation with clip latents". In: *arXiv preprint arXiv:2204.06125* (2022).

[SE19] Yang Song and Stefano Ermon. "Generative Modeling by Estimating Gradients of the Data Distribution". In: *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems*. 2019.

[SE20] Yang Song and Stefano Ermon. "Improved techniques for training score-based generative models". In: *arXiv preprint arXiv:2006.09011* (2020).

[Son+20] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. "Score-Based Generative Modeling through Stochastic Differential Equations". In: *International Conference on Learning Representations*. 2020.

[Son+21a] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. "Maximum likelihood training of score-based diffusion models". In: *Advances in Neural Information Processing Systems* 34 (2021).

[Son+21b] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. "Solving Inverse Problems in Medical Imaging with Score-Based Generative Models". In: *arXiv preprint arXiv:2111.08005* (2021).

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press, 2018.

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

[ZC22] Qinsheng Zhang and Yongxin Chen. "Fast Sampling of Diffusion Models with Exponential Integrator". In: *arXiv preprint arXiv:2204.13902* (2022).

Notation and definitions

We let $\widetilde{p}_t$ denote the density of $\widetilde{x}_t$ under the forward process (1). Note that $x_0 \sim \widetilde{P}_0$ may not admit a density, but $\widetilde{x}_t$ will for $t > 0$. For the reverse process, we use the notation $p_t = \widetilde{p}_{T-t}$, $x_t = \widetilde{x}_{T-t}$. We defined m_t and σ_t in (3),

$$m_t = \exp\left[-\frac{1}{2}\int_0^t g(s)^2\,ds\right], \qquad \sigma_t^2 = 1 - \exp\left[-\int_0^t g(s)^2\,ds\right],$$

and note that $\widetilde{p}_t = (M_{m_t\sharp}\widetilde{P}_0) * \varphi_{\sigma_t^2}$, where $M_m(x) = mx$ denotes multiplication by m, $F_\sharp P$ denotes the pushforward of the measure P by F, and φ_{σ^2} is the density of $N(0, \sigma^2 I_d)$. When $g \equiv 1$, we note the bound $\sigma_t^2 \le \min\{1, t\}$ and $\sigma_t^2 = \Theta(\min\{1, t\})$.

We will let z_t denote the (interpolated) discrete process (see (12)) and let q_t be the density of z_t. We define

$$\phi_t(x) = \frac{q_t(x)}{p_t(x)}, \qquad \psi_t(x) = \frac{\phi_t(x)}{\mathbb{E}_{p_t}\phi_t^2}, \tag{8}$$

and note that $q_t\psi_t$ is a probability density. We defined $G_{t',t} = \int_{t'}^t g(T-s)^2\,ds$ in (6).

We denote the estimated score function by either $s(x,t)$ and $s_t(x)$ interchangeably.

A random variable X is subgaussian with constant C if

$$C = \|X\|_{\psi_2} := \inf\left\{t > 0 : \mathbb{E}\exp(X^2/t^2) \le 2\right\} < \infty.$$

A $\mathbb{R}^d$-valued random variable X is subgaussian with constant C if for all $v \in \mathbb{S}^{d-1}$, $\langle X, v\rangle$ is subgaussian. We also define

$$\|X\|_{2,\psi_2} := \left\|\|X\|_2\right\|_{\psi_2}.$$

Given a probability measure P on $\mathbb{R}^d$ with density p, the associated Dirichlet form is $\mathscr{E}_p(f,g) := \int_{\mathbb{R}^d}\langle\nabla f, \nabla g\rangle\,P(dx) = \int_{\mathbb{R}^d}\langle\nabla f, \nabla g\rangle\,p(x)\,dx$; denote $\mathscr{E}_p(f) = \mathscr{E}_p(f,f)$. we say that a log-Sobolev inequality (LSI) holds with constant C_{LS} if for any probability density q,

$$\mathrm{KL}(q||p) \le \frac{C_{\text{LS}}}{2}\mathscr{E}_p\left(\frac{q}{p}, \ln\frac{q}{p}\right) = \frac{C_{\text{LS}}}{2}\int_{\mathbb{R}^d}\left\|\nabla\ln\frac{q(x)}{p(x)}\right\|^2 q(x)\,dx. \tag{9}$$

Note $\int_{\mathbb{R}^d}\left\|\nabla\ln\frac{q(x)}{p(x)}\right\|^2 q(x)\,dx$ is also known as the Fisher information of q with respect to p. Alternatively, defining the entropy by $\mathrm{Ent}_p(f) = \mathbb{E}_p f(x)\ln f(x) - \mathbb{E}_p f(x)\ln\mathbb{E}_p f(x)$, for any $f \ge 0$,

$$\mathrm{Ent}_p(f) \le \frac{C_{\text{LS}}}{2}\mathscr{E}_p(f, \ln f) = \frac{C_{\text{LS}}}{2}\int_{\mathbb{R}^d}\|\nabla\ln f(x)\|^2 f(x)p(x)dx. \tag{10}$$

A Discussion

A.1 Prior work on convergence guarantees

We highlight two recent works which make progress on this problem. [LLT22] are the first to give polynomial convergence guarantees in TV distance under L^2-accurate score for a reasonable family

of distributions. They introduce a framework to reduce the analysis under L^2-accurate score to L^∞-accurate score. However, they rely on the data distribution satisfying smoothness conditions and a log-Sobolev inequality—a strong assumption which essentially limits the guarantees to unimodal distributions.

[De 22] instead make minimal data assumptions, giving convergence in Wasserstein distance for distributions with bounded support $\mathcal{M}$. However, their guarantees have dependence $\exp(O(\operatorname{diam}(\mathcal{M})^2/\varepsilon))$ for general distributions, and an improved, but still exponential dependence $\exp(\widetilde{O}(\Gamma))$ for distributions with Hessian bounded by Γ/σ_t^2.

We note that other works also analyze the generalization error of a learned score estimate [BMR20; De 22]. This is an important question because without further assumptions, learning an L^2-accurate score estimate requires a number of samples exponential in the dimension. As this is beyond the scope of our paper, we assume that an L^2-accurate score estimate is obtainable.

A.2 Our contributions

In this work, we analyze convergence in the most general setting of distributions of bounded support, as in [De 22]. We give Wasserstein bounds for *any* distribution of bounded support (or sufficiently decaying tails), and TV bounds for distributions under smoothness assumptions, that are polynomial in all parameters, and do not rely on the data distribution satisfying any functional inequality.

We streamline the χ^2-based analysis of [LLT22], with significant changes as to completely remove the use of functional inequalities. In particular, the biggest challenge—and our key improvement—is to bound a certain KL-divergence without reliance on a global functional inequality. For this, we prove a key lemma that distributions which are close in χ^2-divergence have score functions that are close in L^2, which may be of independent interest.

Upon finishing our paper, we learned of a concurrent and independent work [Che+22] which obtained theoretical guarantees for score-based generative modeling under similarly general assumptions on the data distribution. We note that although our bounds are obtained under similar assumptions (with our assumption of the score estimate accuracy slightly weaker than theirs), our proof techniques are quite different. Following the "bad set" idea from [LLT22], we derived a change-of-measure inequality with Theorem E.1, while the analysis in [Che+22] is based on the Girsanov approach.

B DDPM with L^∞-accurate score estimate

We consider the error between the exact backwards SDE (4) and the exponential integrator with estimated score (5). In this section, we bound the error assuming that the score estimate s is accurate in L^∞.

Assumption 5 (L^∞ score error)**.** *For any $t \in \{T - t_0, \dots, T - t_K\}$, the error in the score estimate is bounded:*

$$\|\nabla \ln \widetilde{p}_t - s(\cdot, t)\|_\infty = \sup_{x \in \mathbb{R}^d} \|\nabla \ln \widetilde{p}_t(x) - s(x,t)\| \le \varepsilon_{\infty,t}^2 \tag{11}$$

for some non-decreasing function $\varepsilon_{\infty,t}^2$.

In Section E, we will relax this condition to score function being accurate in L^2.

First, we construct the following continuous-time process which interpolates the discrete-time process (5), for $t \in [t_k, t_{k+1}]$:

$$dz_t = g(T-t)^2 \left(\frac{1}{2} z_t + s(z_{t_k}, T - t_k)\right) dt + g(T-t)\, dw_t. \tag{12}$$

Integration gives that

$$\begin{aligned} z_t - z_{t_k} = &\left(\exp\left(\frac{1}{2} G_{t_k,t}\right) - 1\right)(z_{t_k} + 2s(z_{t_k}, T - t_k)) \\ &+ \int_{t_k}^t \exp\left(\frac{1}{2}\int_{t_k}^{t'} g(T-t'')^2\, dt''\right) g(t')\, dw_{t'}, \end{aligned} \tag{13}$$

where $G_{t',t}$ is defined in (6).

Letting q_t be the distribution of z_t and p_t be the distribution of x_t, we have by [LLT22, Lemma A.2] that

$$\frac{d}{dt}\chi^2(q_t||p_t) = -g(T-t)^2\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 2\mathbb{E}\left[\left\langle g(T-t)^2(s(z_{t_k}, T-t_k) - \nabla \ln p_t(z_t), \nabla \frac{q_t(x)}{p_t(x)}\right\rangle\right]. \tag{14}$$

(Note that in our case, $\widehat{f}$ also depends on z_t rather than just z_{t_k}, but this does not change the calculation.) Define ϕ_t, ψ_t as in (8): $\phi_t(x) = \frac{q_t(x)}{p_t(x)}$, $\psi_t(x) = \frac{\phi_t(x)}{\mathbb{E}_{p_t}\phi_t^2}$.

To bound (14), we use the following lemma.

Lemma B.1 (cf. [EHZ21, Lemma 1], [LLT22, Lemma A.3])**.** *For any $C > 0$ and any $\mathbb{R}^d$-valued random variable u, we have*

$$\mathbb{E}\left[\left\langle u, \nabla \frac{q_t(z_t)}{p_t(z_t)}\right\rangle\right] \le C \cdot (\chi^2(q_t||p_t) + 1) \cdot \mathbb{E}\left[\|u\|^2 \psi_t(z_t)\right] + \frac{1}{4C}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right).$$

Proof. By Young's inequality,

$$\begin{aligned}
\mathbb{E}\left[\left\langle u, \nabla \frac{q_t(z_t)}{p_t(z_t)}\right\rangle\right] &= \mathbb{E}\left[\left\langle u\sqrt{\frac{q_t(z_t)}{p_t(z_t)}}, \sqrt{\frac{p_t(z_t)}{q_t(z_t)}}\nabla \frac{q_t(z_t)}{p_t(z_t)}\right\rangle\right] \\
&\le C\mathbb{E}\left[\|u\|^2 \frac{q_t(z_t)}{p_t(z_t)}\right] + \frac{1}{4C}\mathbb{E}_{p_t}\left[\left\|\nabla \frac{q_t(x)}{p_t(x)}\right\|^2\right] \\
&= C\mathbb{E}_{p_t}\phi_t^2 \cdot \mathbb{E}\left[\|u\|^2 \psi_t(z_t)\right] + \frac{1}{4C}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) \\
&= C(\chi^2(q_t||p_t) + 1) \cdot \mathbb{E}\left[\|u\|^2 \psi_t(z_t)\right] + \frac{1}{4C}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right). \qquad \square
\end{aligned}$$

Lemma B.2. *Suppose that* (11) *holds for $t = T - t_k$, $\nabla \ln p_{t_k}(x)$ is L_{T-t_k}-Lipschitz, g is non-decreasing, and that $h_k \le \frac{1}{20 L_{T-t_k} g(T-t_k)^2}$. Then we have for $t \in [t_k, t_{k+1}]$ that*

$$\begin{aligned}
\frac{d}{dt}\chi^2(q_t||p_t) \le -\frac{1}{2}g(T-t)^2\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 12g(T-t)^2(\chi^2(q_t||p_t)+1)\cdot \\
\Bigg[\varepsilon_{\infty,T-t_k}^2 + 16G_{t_k,t}^2 L_{T-t_k}^2\left[\mathbb{E}[\psi_t(z_t)\|z_t\|^2] + 16\mathbb{E}[\psi_t(z_t)\|\nabla \ln p_t(x_t)\|^2]\right] \\
+ 64G_{t_k,t}L_{T-t_k}^2(8\,\mathrm{KL}(\psi_t q_t||p_t) + 2d + 16\ln 2) + \mathbb{E}\left[\|\nabla \ln p_{t_k}(z_t) - \nabla \ln p_t(z_t)\|^2 \psi_t(z_t)\right]\Bigg].
\end{aligned}$$

Proof. We bound the second term on the RHS of (14). By Lemma B.1,

$$\begin{aligned}
&\mathbb{E}\left[\left\langle s(z_{t_k}, T-t_k) - \nabla \ln p_t(z_t), \nabla \frac{q_t(z_t)}{p_t(z_t)}\right\rangle\right] \\
&\le (\chi^2(q_t||p_t)+1)\mathbb{E}\left[\|s(z_{t_k}, T-t_k) - \nabla \ln p_t(z_t)\|^2 \psi_t(z_t)\right] + \frac{1}{4}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right).
\end{aligned} \tag{15}$$

Now

$$\begin{aligned}
&\|s(z_{t_k}, T-t_k) - \nabla \ln p_t(z_t)\|^2 \\
&\le 3\left[\|s(z_{t_k}, T-t_k) - \nabla \ln p_{t_k}(z_{t_k})\|^2 + \|\nabla \ln p_{t_k}(z_{t_k}) - \nabla \ln p_{t_k}(z_t)\|^2 + \|\nabla \ln p_{t_k}(z_t) - \nabla \ln p_t(z_t)\|^2\right] \\
&\le 3\left[\|s(z_{t_k}, T-t_k) - \nabla \ln p_{t_k}(z_{t_k})\|^2 + L_{T-t_k}^2\|z_{t_k} - z_t\|^2 + \|\nabla \ln p_{t_k}(z_t) - \nabla \ln p_t(z_t)\|^2\right]
\end{aligned}$$

and

$$\mathbb{E}\left[\|s(z_{t_k}, T-t_k) - \nabla \ln p_{t_k}(z_{t_k})\|^2 \psi_t(z_t)\right] \le \varepsilon^2_{\infty, T-t_k}$$

by definition of $\varepsilon_{\infty,t}$, so by Lemma B.3,

$$\begin{aligned}
&\mathbb{E}\left[\|s(z_{t_k}, T-t_k) - \nabla \ln p_t(z_t)\|^2 \psi_t(z_t)\right] \\
&\le 3\left[\varepsilon^2_{\infty,T-t_k} + L^2_{T-t_k}\mathbb{E}\left[\|z_t - z_{t_k}\|^2 \psi_t(z_t)\right] + \mathbb{E}\left[\|\nabla \ln p_{t_k}(z_t) - \nabla \ln p_t(z_t)\|^2 \psi_t(z_t)\right]\right] \\
&\le 3\Bigg[\varepsilon^2_{\infty,T-t_k} + 16G^2_{t_k,t}L^2_{T-t_k}\Big[\mathbb{E}[\psi_t(z_t)\|z_t\|^2] + 4\mathbb{E}[\psi_t(z_t)\|s(z_{t_k}, T-t_k) - \nabla \ln p_t(z_t)\|^2] \\
&\quad + 16\mathbb{E}[\psi_t(z_t)\|\nabla \ln p_t(x_t)\|^2]\Big] + 64G_{t_k,t}L^2_{T-t_k}(8\,\mathrm{KL}(\psi_t q_t||p_t) + 2d + 16\ln 2) \\
&\quad + \mathbb{E}\left[\|\nabla \ln p_{t_k}(z_t) - \nabla \ln p_t(z_t)\|^2 \psi_t(z_t)\right]\Bigg]
\end{aligned}$$

The condition on h_k and the fact that g is non-decreasing implies $192G^2_{t_k,t}L^2_{T-t_k} \le \frac{1}{2}$. Rearranging gives

$$\begin{aligned}
&\mathbb{E}\left[\|s(z_{t_k}, T-t_k) - \nabla \ln p_t(z_t)\|^2 \psi_t(z_t)\right] \\
&\le 6\Bigg[\varepsilon^2_{\infty,T-t_k} + 16G^2_{t_k,t}L^2_{T-t_k}\left[\mathbb{E}[\psi_t(z_t)\|z_t\|^2] + 16\mathbb{E}[\psi_t(z_t)\|\nabla \ln p_t(x_t)\|^2\right] \\
&\quad + 64G_{t_k,t}L^2_{T-t_k}(8\,\mathrm{KL}(\psi_t q_t||p_t) + 2d + 16\ln 2) + \mathbb{E}\left[\|\nabla \ln p_{t_k}(z_t) - \nabla \ln p_t(z_t)\|^2 \psi_t(z_t)\right]\Bigg]
\end{aligned}$$

Substituting into (15) and that inequality into (14) give the conclusion. □

Lemma B.3. *Suppose that* $h_k \le \frac{1}{2g(T-t_k)^2}$. *Then for* $t \in [t_{k+1}, t_k]$,

$$\begin{aligned}
\mathbb{E}\left[\|z_t - z_{t_k}\|^2 \psi_t(z_t)\right] &\le 16G^2_{t_k,t}\Big[\mathbb{E}[\psi_t(z_t)\|z_t\|^2] + 4\mathbb{E}[\psi_t(z_t)\|s(z_{t_k}, T-t_k) - \nabla \ln p_t(z_t)\|^2] \\
&\quad + 16\mathbb{E}[\psi_t(z_t)\|\nabla \ln p_t(z_t)\|^2]\Big] + 64G_{t_k,t}(8\,\mathrm{KL}(\psi_t q_t||p_t) + 2d + 16\ln 2).
\end{aligned}$$

Proof. Consider (13). The assumption on h_k implies $G_{t_k,t} \le G_{t_k,t_{k+1}} \le \frac{1}{2}$, so $\exp\left(\frac{1}{2}G_{t_k,t}\right) - 1 \le G_{t_k,t}$. Let Y denote the last term of (13). Then

$$\begin{aligned}
\|z_t - z_{t_k}\| &\le G_{t_k,t}\left[\|z_{t_k}\| + 2\|s(z_{t_k}, T-t_k)\|\right] + \|Y\| \\
&\le G_{t_k,t}\left[\|z_t\| + \|z_{t_k} - z_t\| + 2\|s(z_{t_k}, T-t_k) - \nabla \ln p_t(z_t)\| + 2\|\nabla \ln p_t(z_t)\|\right] + \|Y\|.
\end{aligned}$$

Again using $G_{t_k,t} \le \frac{1}{2}$, rearranging gives

$$\|z_t - z_{t_k}\| \le 2G_{t_k,t}\left[\|z_t\| + 2\|s(z_{t_k}, T-t_k) - \nabla \ln p_t(z_t)\| + 4\|\nabla \ln p_t(z_t)\|\right] + 2\|Y\|,$$

and

$$\begin{aligned}
\mathbb{E}\left[\|z_t - z_{t_k}\|^2 \psi_t(z_t)\right] &\le 16G^2_{t_k,t}\Big[\mathbb{E}[\psi_t(z_t)\|z_t\|^2] + 4\mathbb{E}[\psi_t(z_t)\|s(z_{t_k}, T-t_k) - \nabla \ln p_t(z_t)\|^2] \\
&\quad + 16\mathbb{E}[\psi_t(z_t)\|\nabla \ln p_t(x_t)\|^2]\Big] + 16\mathbb{E}[\psi_t(z_t)\|Y\|^2].
\end{aligned}$$

By Lemma B.4,

$$\mathbb{E}[\psi_t(z_t)\|Y\|^2] \le 4G_{t_k,t}(8\,\mathrm{KL}(\psi_t q_t||p_t) + 2d + 16\ln 2).$$

The lemma follows. □

Lemma B.4. *For* $t \in [t_k, t_{k+1}]$,

$$\begin{aligned}
&\mathbb{E}\left[\psi_t(z_t)\left\|\int_{t_k}^{t} \exp\left(\frac{1}{2}\int_{t_k}^{t'} g(T-t'')^2\,dt''\right) g(t')\,dw_{t'}\right\|^2\right] \\
&\quad \le 2(\exp(G_{t_k,t}) - 1)\big(8\,\mathrm{KL}(\psi_t q_t||p_t) + 2d + 16\ln 2\big).
\end{aligned}$$

Proof. Note that $Y := \int_{t_k}^t \exp\left(\frac{1}{2}\int_{t_k}^{t'} g(T-t'')^2\,dt''\right) g(t')\,dw_{t'}$ is a Gaussian random vector with variance

$$\int_{t_k}^t \exp\left(\int_{t_k}^{t'} g(T-t'')^2\,dt''\right) g(t')^2\,dt' \cdot I_d = \exp\left(\int_{t_k}^{t'} g(T-t'')^2\,dt''\right)\Big|_{t'=t_k}^{t'=t} \cdot I_d$$
$$= (\exp(G_{t_k,t}) - 1) \cdot I_d.$$

(Note that this calculation shows that the continuous-time process (12) does agree with the discrete-time process (5) at $t = t_{k+1}$.) Using the Donsker-Varadhan variational principle, for any random variable X,

$$\tilde{\mathbb{E}} X \le \mathrm{KL}(\tilde{\mathbb{P}}||\mathbb{P}) + \ln \mathbb{E} \exp X.$$

Applying this to $X = c\left(\|Y\| - \mathbb{E}\|Y\|\right)^2$ for a constant $c > 0$ to be chosen later, and $\widetilde{\mathbb{P}}$ such that $\frac{d\widetilde{\mathbb{P}}}{d\mathbb{P}}(z_t) = \psi_t(z_t)$, we can bound

$$\begin{aligned}
\tilde{\mathbb{E}}\|Y\|^2 &\le 2\mathbb{E}\left[\|Y\|^2\right] + 2\widetilde{\mathbb{E}}\left[(Y - \mathbb{E}\|Y\|)^2\right] \\
&\le 2\mathbb{E}\left[\|Y\|^2\right] + \frac{2}{c}\left[\mathrm{KL}(\tilde{\mathbb{P}}||\mathbb{P}) + \ln \mathbb{E}\exp\left(c\left(\|Y\| - \mathbb{E}\|Y\|\right)^2\right)\right] \qquad (16) \\
&\le 2d(\exp(G_{t_k,t}) - 1) + \frac{2}{c}\left[\mathrm{KL}(\tilde{\mathbb{P}}||\mathbb{P}) + \ln \mathbb{E}\exp\left(c\left(\|Y\| - \mathbb{E}\|Y\|\right)^2\right)\right]. \qquad (17)
\end{aligned}$$

Now following [Che+21, Theorem 4], we set $c = \frac{1}{8(\exp(G_{t_k,t})-1)}$, so that

$$\mathbb{E}\left[\frac{(\|Y\| - \mathbb{E}\|Y\|)^2}{8(\exp(G_{t_k,t}) - 1)}\right] \le 2.$$

Next, we have

$$\begin{aligned}
\mathrm{KL}(\tilde{\mathbb{P}}||\mathbb{P}) &= \mathbb{E}_{\psi_t q_t} \ln \psi_t = \mathbb{E}_{\psi_t q_t} \ln \frac{\phi_t}{\mathbb{E}_{p_t}\phi_t^2} = \frac{1}{2}\mathbb{E}_{\psi_t q_t} \ln \frac{\phi_t^2}{(\mathbb{E}_{p_t}\phi_t^2)^2} \\
&= \frac{1}{2}\left[\mathbb{E}_{\psi_t q_t} \ln \frac{\phi_t^2}{\mathbb{E}_{p_t}\phi_t^2} - \ln \mathbb{E}_{p_t}\phi_t^2\right] = \frac{1}{2}\left[\mathbb{E}_{\psi_t q_t} \ln \frac{\psi_t q_t}{p_t} - \ln \mathbb{E}_{p_t}\phi_t^2\right].
\end{aligned}$$

Noting that $\mathbb{E}_{p_t}\phi_t^2 = \chi^2(q_t||p_t) + 1 \ge 1$, we have that

$$\mathrm{KL}(\tilde{\mathbb{P}}||\mathbb{P}) \le \frac{1}{2}\mathrm{KL}(\psi_t q_t||p_t).$$

Substituting everything into (17) gives the desired inequality. □

Let

$$K_z = \mathbb{E}\left[\psi_t(z_t)\|z_t\|^2\right] \qquad (18)$$
$$K_V = \mathbb{E}\left[\psi_t(z_t)\|\nabla \ln p_t(z_t)\|^2\right] \qquad (19)$$
$$K_{\Delta V} = \mathbb{E}\left[\psi_t(z_t)\|\nabla \ln p_{t_k}(z_t) - \nabla \ln p_t(z_t)\|^2\right] \qquad (20)$$
$$K = \mathrm{KL}(\psi_t q_t||p_t). \qquad (21)$$

In order to bound the RHS in Lemma B.2, we need to bound all four of these quantities, which we do in Lemma B.5, B.6, B.8, and Section C, respectively. The main innovation in our analysis compared to [LLT22] is a new way to bound K, which we present in a separate section.

First we bound K_z. Recall the norm

$$\|X\|_{2,\psi_2} = \inf\left\{L > 0 : \mathbb{E}e^{\frac{\|X\|_2^2}{L^2}} \le 2\right\}.$$

(In other words, this is the usual Orlicz norm applied to $\|X\|_2$.)

Lemma B.5. *For $t \in [t_k, t_{k+1}]$,*

$$\mathbb{E}\left[\psi_t(z_t)\left\|z_t\right\|^2\right] \le \left\|x_t\right\|_{2,\psi_2}^2 \cdot \left[\mathrm{KL}(\psi_t q_t || p_t) + \ln 2\right].$$

Proof. By the Donsker-Varadhan variational principle,

$$\mathbb{E}\left[\psi_t(z_t)\left\|z_t\right\|^2\right] = \frac{2}{s}\mathbb{E}_{\psi_t q_t}\left[\frac{s}{2}\left\|x\right\|^2\right] \le \frac{2}{s}\left[\mathrm{KL}(\psi_t q_t || p_t) + \ln \mathbb{E}_{p_t}\left[e^{\frac{s}{2}\|x\|^2}\right]\right]$$

for any $s > 0$. Choosing $s = 2\left\|x_t\right\|_{2,\psi_2}^{-2}$, we have $\mathbb{E}_{p_t}\left[e^{\frac{s}{2}\|x\|^2}\right] \le 2$, which gives the desired inequality. □

The following bounds K_V; note that the proof does not depend on the definition of q_t, only that it is a probability density.

Lemma B.6 ([LLT22, Corollary C.7], [Che+21, Lemma 16])**.**

$$\mathbb{E}\left[\psi_t(z_t)\left\|\nabla \ln p_t(z_t)\right\|^2\right] \le \frac{4}{\chi^2(q_t||p_t)+1} \cdot \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 2dL.$$

We use the following lemma to bound $K_{\Delta V}$ in Lemma B.8.

Lemma B.7 ([LLT22, Lemma C.12])**.** *Suppose that $p(x) \propto e^{-V(x)}$ is a probability density on $\mathbb{R}^d$, where $V(x)$ is L-smooth. Let $p_\alpha(x) = \alpha^d p(\alpha x)$ and $\varphi_{\sigma^2}(x)$ denote the density function of $N(0, \sigma^2 I_d)$. Then for $\sigma^2 \le \frac{1}{2\alpha^2 L}$,*

$$\left\|\nabla \ln \frac{p(x)}{(p_\alpha * \varphi_{\sigma^2})(x)}\right\| \le 6\alpha^2 L\sigma d^{1/2} + (\alpha + 2\alpha^3 L\sigma^2)(\alpha-1)L\left\|x\right\| + (\alpha - 1 + 2\alpha^3 L\sigma^2)\left\|\nabla V(x)\right\|.$$

Lemma B.8. *Suppose that $h_k \le \frac{1}{4Lg(T-t_k)^2}$ where $\nabla \ln p_t$ is L_{T-t}-smooth ($L_{T-t} \ge 1$) and $L = \max_{t\in[t_k,t_{k+1}]} L_{T-t}$. For $t \in [t_k, t_{k+1}]$,*

$$\begin{aligned}
&\mathbb{E}\left[\psi_t(z_t)\left\|\nabla \ln p_{t_k}(z_t) - \nabla \ln p_t(z_t)\right\|^2\right] \\
&\le 25L_{T-t}^2\left(8G_{t_k,t}d + G_{t_k,t}^2\mathbb{E}\left[\psi_t(z_t)\left\|z_t\right\|^2\right]\right) + 100L_{T-t}^2 G_{t_k,t}^2 \mathbb{E}\left[\psi_t(z_t)\left\|\nabla \ln p_t(z_t)\right\|^2\right]
\end{aligned}$$

Proof. We have the following relationship for $t \in [t_k, t_{k+1}]$:

$$p_{t_k} = (p_t)_\alpha * \varphi_{\sigma^2}.$$

where $p_\alpha(x) = \alpha^d p(\alpha x)$, $\alpha = e^{\frac{1}{2}\int_{t_k}^t g(T-s)^2\,ds}$ and $\sigma^2 = 1 - e^{-\int_{t_k}^t g(T-s)^2\,ds}$. Observe that since $h_k \le \frac{1}{4g(T-t_k)^2}$,

$$\alpha \le 1 + \int_{t_k}^t g(T-s)^2 ds \le 1 + h_k g(T-t_k)^2 \le 1 + \frac{1}{4}$$

$$\sigma^2 = 1 - e^{-\int_{t_k}^t g(T-s)^2 ds} \le \int_{t_k}^t g(T-s)^2 ds \le h_k g(T-t_k)^2 \le \frac{1}{4}.$$

We note that

$$\sigma^2 \le h_k g(T-t_k)^2 \le \frac{1}{4L_t} \le \frac{1}{2\alpha^2 L_t}$$

so the hypothesis of Lemma B.7 is satisfied. Using Lemma B.7, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\psi_t(z_t)\left\|\nabla \ln p_{t_k}(z_t) - \nabla \ln p_t(z_t)\right\|^2\right] \\
&\le 72\alpha^4 L_{T-t}^2 \sigma^2 d + 4(\alpha + 2\alpha^3 L_{T-t}\sigma^2)^2(\alpha-1)^2 L_{T-t}^2 \mathbb{E}\left[\psi(z_t)\left\|z_t\right\|^2\right] \\
&\quad + 4(\alpha - 1 + 2\alpha^3 L_{T-t}\sigma^2)^2 \mathbb{E}\left[\psi_t(z_t)\left\|\nabla \ln p_t(z_t)\right\|^2\right] \\
&\le 72(5/4)^4 L_{T-t}^2 G_{t_k,t} d + 4(2\alpha)^2 G_{t_k,t}^2 L_{T-t}^2 \mathbb{E}\left[\psi(z_t)\left\|z_t\right\|^2\right] \\
&\quad + 4(G_{t_k,t} + 4L_{T-t}G_{t_k,t})^2 \mathbb{E}\left[\psi_t(z_t)\left\|\nabla \ln p_t(z_t)\right\|^2\right] \\
&\le 200L_{T-t}^2 d G_{t_k,t} + 25L_{T-t}^2 G_{t_k,t}^2 \mathbb{E}\left[\psi(z_t)\left\|z_t\right\|^2\right] + 100L_{T-t}^2 G_{t_k,t}^2 \mathbb{E}\left[\psi_t(z_t)\left\|\nabla \ln p_t(z_t)\right\|^2\right].
\end{aligned}$$

$\square$

Now we put everything together. Write $G_t = G_{t_k,t}$ for short. Suppose L_t is non-increasing. By Lemma B.2,

$$\frac{d}{dt}\chi^2(q_t||p_t) \le -\frac{1}{2}g(T-t)^2\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 12g(T-t)^2(\chi^2(q_t||p_t)+1)\cdot E$$

where $E \le 16G_t^2L_{T-t_k}^2(K_z+16K_V) + 64G_tL_{T-t_k}^2(8K+2d+16\ln 2) + \varepsilon_{\infty,T-t_k}^2 + K_{\Delta V}$.

By Lemma B.8, $K_{\Delta V} \le 25L_{T-t}^2(8G_td + G_t^2K_z) + 100L_{T-t}^2G_t^2K_V$, so

$$E \le 41G_t^2L_{T-t}^2K_z + 356G_t^2L_{T-t}^2K_V + 64G_tL_{T-t}^2(8K+6d+16\ln 2) + \varepsilon_{\infty,T-t_k}^2.$$

By Lemma B.5, $K_z \le \|x_t\|_{2,\psi_2}^2(K+\ln 2)$, and by Corollary B.6, $K_V \le \frac{4}{\chi^2(q_t||p_t)+1}\cdot\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 2dL$, so

$$\begin{aligned}E &\le 41G_t^2L_{T-t}^2\left(\|x_t\|_{2,\psi_2}^2(K+\ln 2)\right) + 356G_t^2L_{T-t}^2\left(\frac{4}{\chi^2(q_t||p_t)+1}\cdot\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right)+2dL\right)\\ &\quad + 64G_tL_{T-t}^2(8K+6d+16\ln 2) + \varepsilon_{\infty,T-t_k}^2.\end{aligned}$$

Now, if $h_k \le \frac{\varepsilon'_{h_k}}{20g(T-t_k)^2L_{T-t_{k+1}}}$, then

$$\begin{aligned}E &\le {\varepsilon'_{h_k}}^2\left[\|x_t\|_{2,\psi_2}^2(K+\ln 2) + \left(\frac{4}{\chi^2(q_t||p_t)+1}\cdot\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right)+2dL_{T-t}\right)\right]\\ &\quad + 4\varepsilon'_{h_k}L_{T-t}(8K+2d+16\ln 2) + \varepsilon_{\infty,T-t_k}^2.\end{aligned}$$

Let $M_{T-t} := \|x_t\|_{2,\psi_2}^2$. Assume that $K \le \frac{A_{T-t}}{\chi^2(q_t||p_t)+1} + B_{T-t}$. Then we obtain

$$\begin{aligned}&12g(T-t)^2(\chi^2(q_t||p_t)+1)\cdot E\\ &\le 12g(T-t)^2\Big[\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right)({\varepsilon'_{h_k}}^2\cdot(A_{T-t}M_{T-t}+4) + \varepsilon'_{h_k}\cdot 32L_{T-t}A_{T-t})\\ &\qquad + (\chi^2(q_t||p_t)+1)({\varepsilon'_{h_k}}^2\cdot((B_{T-t}+\ln 2)M_{T-t}+2dL)\\ &\qquad + \varepsilon'_{h_k}\cdot L_{T-t}(8B_{T-t}+6d+16\ln 2)) + \varepsilon_{\infty,T-t_k}^2\Big].\end{aligned}$$

If $\varepsilon'_{h_k} \le \min\left\{\frac{1}{\sqrt{48(A_{T-t}M_{T-t}+4)}}, \frac{1}{128L_{T-t}A_{T-t}}\right\}$, then

$$\begin{aligned}\frac{d}{dt}\chi^2(q_t||p_t) &\le 12g(T-t)^2\Big[(\chi^2(q_t||p_t)+1)({\varepsilon'_{h_k}}^2\cdot((B_{T-t}+\ln 2)M_{T-t}+2dL_{T-t})\\ &\qquad + \varepsilon'_{h_k}\cdot L_{T-t}(8B_{T-t}+6d+16\ln 2)) + \varepsilon_{\infty,T-t_k}^2\Big].\end{aligned}$$

If $\varepsilon'_{h_k} \le \min\left\{\frac{\sqrt{\varepsilon'}}{g(T-t)\sqrt{24(T-t_k)((B_{T-t}+\ln 2)M_{T-t}+2dL_{T-t})}}, \frac{\varepsilon'}{24g(T-t)^2(T-t)L_{T-t}(8B_{T-t}+6d+16\ln 2)}\right\}$, we get

$$\frac{d}{dt}\chi^2(q_t||p_t) \le \frac{\varepsilon'}{T-t}(\chi^2(q_t||p_t)+1) + \varepsilon_{\infty,T-t_k}^2g(T-t)^2.$$

Integration gives

$$\begin{aligned}\chi^2(q_{t_k}||p_{t_k}) &\le e^{\varepsilon'\int_0^{t_k}\frac{1}{T-t}\,dt}(\chi^2(q_0||p_0)+1) + \int_0^{t_k} e^{\int_t^{t_k}\frac{\varepsilon'}{T-s}\,ds}\varepsilon_{T-t}^2g(T-t)^2\,dt\\ &\le \left(\frac{T}{T-t_k}\right)^{\varepsilon'}\chi^2(q_0||p_0) + \left(\left(\frac{T}{T-t_k}\right)^{\varepsilon'}-1\right) + \int_0^{t_k}\left(\frac{T-t}{T-t_k}\right)^{\varepsilon'}\varepsilon_{T-t}^2g(T-t)^2\,dt.\end{aligned}$$

Taking $\varepsilon' = \frac{\varepsilon}{\ln\left(\frac{T}{T-t_N}\right)}$ then gives the following Theorem B.10. We first introduce a technical assumption.

Definition B.9. *Let $f : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$. We say that f has at most power growth and decay (with some constant $c > 0$) if $\max_{u \in [\frac{t}{2}, t]} f(u) \in \left[\frac{f(t)}{c}, cf(t)\right]$.*

Theorem B.10. *Suppose that the following hold.*

1. *Assumption 5 holds for $\varepsilon_{\infty,t}$.*
2. $\|\widetilde{x}_t\|_{2,\psi_2}^2 \leq M_t$.
3. *The KL bound $\mathrm{KL}(\psi_t q_t || p_t) \leq \frac{A_{T-t}}{\chi^2(q_t||p_t)+1} + B_{T-t}$ holds for any density q_t and $t < t_N$, where $\psi_t(x) = \frac{q_t(x)/p_t(x)}{\chi^2(q_t||p_t)+1}$.*
4. *$g(t), A_t, B_t, L_t, M_t$ have at most polynomial growth and decay (with some constant c).*

Then there is some constant c' (depending on c) such that if the step sizes satisfy

$$h_k \leq \min\left\{\frac{T-t_k}{2}, \frac{c'\varepsilon'_{h_k}}{g(T-t_k)^2 L_{T-t_k}}\right\},$$

$$\text{where } \varepsilon'_{h_k} = \min\left\{\frac{1}{\sqrt{A_{T-t_k}M_{T-t_k}+1}}, \frac{1}{L_{T-t_k}A_{T-t_k}}, \frac{\sqrt{\varepsilon/\ln\left(\frac{T}{T-t_N}\right)}}{g(T-t_k)\sqrt{(T-t_k)((B_{T-t_k}+1)M_{T-t_k}+dL_{T-t_k})}}, \frac{\varepsilon/\ln\left(\frac{T}{T-t_N}\right)}{g(T-t_k)^2(T-t_k)L_{T-t_k}(B_{T-t_k}+d)}\right\},$$

then for $0 \leq k \leq N$,

$$\chi^2(q_{t_k}||p_{t_k}) \leq e^{\varepsilon}\chi^2(q_0||p_0) + (e^{\varepsilon}-1) + e^{\varepsilon}\int_0^{t_k} \varepsilon_{\infty,T-t}^2 g(T-t)^2\,dt.$$

Proof. This follows from the above calculations and the observation that if we replace $F(T-t)$ by $F(T-t_k)$, for some F satisfying the power growth and decay assumption, then we change the bound by at most a constant factor, because the step size satisfies $h_k = t_{k+1} - t_k \leq \frac{T-t_k}{2}$. □

We specialize this theorem in the case of distributions with bounded support. Note that although not every initial distribution $\widetilde{p}_t$ may satisfy a KL inequality as required by condition 3 of Theorem 30, Lemma C.2 will give the existence of a distribution that does, and is close in TV-error. Later in Section D, we show that this will have a small effect on the score function, and hence allow us to prove our main theorems.

Corollary B.11. *Suppose that Assumptions 5 and 2 hold, $R^2 \geq d$, $g \equiv 1$, and that $\widetilde{P}_0$ is such that the KL inequality (30) holds. Let $\delta = T - t_N$. If $0 < \delta, \varepsilon < \frac{1}{2}$, $h_k = O\left(\frac{\varepsilon}{\max\{T-t_k,(T-t_k)^{-3}\}R^4 d\ln\left(\frac{T}{\delta}\right)\ln\left(\frac{R}{\delta\varepsilon_K}\right)}\right)$, then for any $0 \leq k \leq N$,*

$$\chi^2(q_{t_k}||p_{t_k}) \leq e^{\varepsilon}\chi^2(q_0||p_0) + \varepsilon + e^{\varepsilon}\int_0^{t_k} \varepsilon_{\infty,T-t}^2\,dt.$$

Proof. For $g \equiv 1$, note that $\sigma_{T-t}^2 = \Theta(\min\{T-t, 1\})$. From Lemma B.13, we can choose

$$L_t = \frac{R^2}{\sigma_t^4} = O\left(\frac{R^2}{\min\{(T-t)^2, 1\}}\right).$$

From Lemma B.15, we can choose

$$M_t = \max\{R^2, d\}.$$

The KL inequality (30) gives us

$$A_t = 6(e+1)\sigma_t^2 = O(\min\{T-t, 1\})$$
$$B_t = \ln\left(\frac{1}{\varepsilon}\right) + d\ln\left(1 + O\left(\frac{R}{\sqrt{T-t_N}}\right)\right)$$

We now check the requirements on h_k. We need

$$\varepsilon'_{h_k} = O\left(\frac{1}{\sqrt{A_{T-t_k}M_{T-t_k}+1}}\right) \qquad \Longleftarrow \varepsilon'_{h_k} = O\left(\frac{1}{\max\{R, \sqrt{d}\}}\right) \tag{22}$$

$$\varepsilon'_{h_k} = O\left(\frac{1}{L_{T-t_k}A_{T-t_k}}\right) \qquad \Longleftarrow \varepsilon'_{h_k} = O\left(\frac{T-t_k}{R^2}\right) \tag{23}$$

$$\varepsilon'_{h_k} = O\left(\frac{\sqrt{\varepsilon/\ln\left(\frac{T}{\delta}\right)}}{\sqrt{(T-t_k)((B_{T-t_k}+1)M_{T-t_k}+dL_{T-t_k})}}\right). \tag{24}$$

For $T - t_k \leq 1$, (24) is implied by

$$\varepsilon'_{h_k} = O\left(\frac{\sqrt{\varepsilon/\ln\left(\frac{T}{\delta}\right)}}{\sqrt{(T-t_k)\left(\ln\left(\frac{1}{\varepsilon_K}\right) + d\ln\left(\frac{R}{\delta}\right)\right)\max\{R^2, d\} + \frac{dR^2}{T-t_k}}}\right)$$
$$\Longleftarrow \varepsilon'_{h_k} = O\left(\sqrt{\frac{\varepsilon(T-t_k)}{d\max\{R^2,d\}\ln\left(\frac{T}{\delta}\right)\ln\left(\frac{R}{\delta\varepsilon_K}\right)}}\right),$$

and for $T - t_k > 1$,

$$\varepsilon'_{h_k} = O\left(\frac{\sqrt{\varepsilon/\ln\left(\frac{T}{\delta}\right)}}{\sqrt{T\left(\ln\left(\frac{1}{\varepsilon}\right) + d\ln\left(\frac{R}{\delta}\right)\right)\max\{R^2,d\} + dR^2}}\right)$$
$$\Longleftarrow \varepsilon'_{h_k} = O\left(\sqrt{\frac{\varepsilon}{Td\max\{R^2,d\}\ln\left(\frac{T}{\delta}\right)\ln\left(\frac{R}{\delta\varepsilon_K}\right)}}\right).$$

Finally, the last requirement is

$$\varepsilon'_{h_k} = O\left(\frac{\varepsilon/\ln\left(\frac{T}{\delta}\right)}{(T-t_k)L_{T-t_k}(B_{T-t_k}+d)}\right)$$
$$\Longleftarrow \varepsilon'_{h_k} = O\left(\frac{\varepsilon}{R^2\max\{T-t_k, (T-t_k)^{-1}\}d\ln\left(\frac{T}{\delta}\right)\ln\left(\frac{R}{\delta\varepsilon_K}\right)}\right).$$

As long as $R^2 = \Omega(d)$ and $\varepsilon < 1$, the last equation implies all the others. Plugging this into Theorem B.10 gives the result. □

Above, we use the Hessian bound $\left\|\nabla^2 \ln p_t(x)\right\| \leq \frac{R^2}{\sigma_t^4}$ given in Lemma B.13. Under the stronger smoothness assumption given by Assumption 3, we can take the step sizes to be larger.

Corollary B.12. *Suppose that Assumptions 5, 2, 3 hold, $C \geq R^2 \geq d$, $g \equiv 1$, and that $\widetilde{P}_0$ is such that the KL inequality* (30) *holds. Let $\delta = T - t_N$. If $0 < \delta, ep < \frac{1}{2}$ and $\varepsilon < 1/\sqrt{T}$, $h_k = O\left(\frac{\varepsilon}{\max\{T-t_k,(T-t_k)^{-1}\}C^2 d\ln\left(\frac{T}{\delta}\right)\ln\left(\frac{R}{\delta\varepsilon_K}\right)}\right)$, then for any $0 \leq k \leq N$,*

$$\chi^2(q_{t_k}||p_{t_k}) \leq e^{\varepsilon}\chi^2(q_0||p_0) + \varepsilon + e^{\varepsilon}\int_0^{t_k} \varepsilon^2_{\infty,T-t}\,dt.$$

Proof. We instead have the bound $L_t = \frac{C}{\sigma_t^2}$. The requirement (22) stays the same, while (23) is implied by $\varepsilon'_{h_k} = O(1/C)$. Inequality (24), for $T - t_k \leq 1$, is implied by

$$\varepsilon'_{h_k} = O\left(\sqrt{\frac{1}{d\max\{C, R^2\}\ln\left(\frac{T}{\delta}\right)\ln\left(\frac{R}{\delta\varepsilon_K}\right)}}\right).$$

and for $T - t_k > 1$,

$$\varepsilon'_{h_k} = O\left(\sqrt{\frac{\varepsilon}{Td\max\{C, R^2\}\ln\left(\frac{T}{\delta}\right)\ln\left(\frac{R}{\delta\varepsilon_K}\right)}}\right).$$

Finally, the last requirement is implied by

$$\varepsilon'_{h_k} = O\left(\frac{\varepsilon}{Cd\ln\left(\frac{T}{\delta}\right)\ln\left(\frac{R}{\delta\varepsilon_K}\right)}\right),$$

and for $C \geq R^2$, $\varepsilon \leq 1/\sqrt{T}$, implies all the others. □

B.1 Auxiliary bounds

In this section we give bounds on the Hessian (L_t, Lemma B.13), initial χ^2 divergence $\chi^2(q_0||p_0)$ (Lemma B.14), and Orlicz norm (M_t, Lemma B.15).

Lemma B.13 (Hessian bound)**.** *Suppose that μ is a probability measure supported on a bounded set $\mathcal{M} \subset \mathbb{R}^d$ with radius R. Then letting φ_{σ^2} denote the density of $N(0, \sigma^2 I_d)$,*

$$\left\|\nabla^2 \ln(\mu * \varphi_{\sigma^2}(x))\right\| \leq \max\left\{\frac{R^2}{\sigma^4}, \frac{1}{\sigma^2}\right\}. \tag{25}$$

Therefore, for $\widetilde{P}_0$ supported on $B_R(0)$, $R \geq 1$, we have

$$\left\|\nabla^2 \ln \widetilde{p}_t(x)\right\| \leq \frac{R^2}{\sigma_t^4}. \tag{26}$$

Proof. Let μ_{x,σ^2} denote the density $\mu(du)$ weighted with the gaussian $\varphi_{\sigma^2}(u - x)$, that is, $\mu_{x,\sigma^2}(du) = \frac{e^{-\frac{\|x-u\|^2}{2\sigma^2}}\mu(du)}{\int_{\mathbb{R}^d} e^{-\frac{\|x-u\|^2}{2\sigma^2}}\mu(du)}$. We note the following calculations:

$$\nabla \ln(\mu * \varphi_{\sigma^2}(x)) = \frac{\nabla \int_{\mathbb{R}^d} e^{-\frac{\|x-u\|^2}{2\sigma^2}}\mu(du)}{\int_{\mathbb{R}^d} e^{-\frac{\|x-u\|^2}{2\sigma^2}}\mu(du)} = \frac{\int_{\mathbb{R}^d} -\frac{x-u}{\sigma^2} e^{-\frac{\|x-u\|^2}{2\sigma^2}}\mu(du)}{\int_{\mathbb{R}^d} e^{-\frac{\|x-u\|^2}{2\sigma^2}}\mu(du)} = -\frac{1}{\sigma^2}\mathbb{E}_{\mu_{x,\sigma^2}}(x - u) \tag{27}$$

$$\nabla^2 \ln(\mu * \varphi_{\sigma^2}(x)) = \frac{1}{\sigma^4}\operatorname{Cov}_{\mu_{x,\sigma^2}}(x - u) - \frac{1}{\sigma^2}I_d = \frac{1}{\sigma^4}\operatorname{Cov}_{\mu_{x,\sigma^2}}(u) - \frac{1}{\sigma^2}I_d. \tag{28}$$

The covariance of a distribution supported on a set of radius R is bounded by R^2 in operator norm. Inequality (25) then follows from (28).

For (26), note that $\widetilde{p}_t = M_{m_t\sharp}\widetilde{P}_0 * \varphi_{\sigma_t^2}$, where m_t is given by (3) and M_m denotes multiplication by m. Since $M_{m_t\sharp}\widetilde{P}_0$ is supported on $B_{m_t R}(0) \subset B_R(0)$ and $\sigma_t \leq 1$, the result follows. □

Lemma B.14 (Bound on initial χ^2-divergence)**.** *Suppose that $\widetilde{P}_0$ is supported on $B_R(0)$. Let $p_{\text{prior}} = N(0, (1 - e^{G_{0,t}})I_d)$. Then*

$$\chi^2(p_{\text{prior}}||\widetilde{p}_T) \leq \exp\left[\frac{R^2\exp(-G_{0,T})}{1 - \exp(-G_{0,T})}\right]$$

and for $0 < \varepsilon < \frac{1}{2}$ and $G_{0,T} \geq \ln\left(\frac{4R^2}{\varepsilon^2}\right) \vee 1$, we have $\chi^2(p_{\text{prior}}||\widetilde{p}_T) \leq \varepsilon^2$.

Proof. We have for $x_0 \sim \widetilde{P}_0$ that

$$\chi^2\left(N(0,(1-e^{-G_{0,T}})I_d)||N(x_0\exp\left(-\frac{1}{2}G_{0,T}\right),(1-\exp(-G_{0,T}))I_d)\right)$$
$$\leq \exp\left[\frac{\|x_0\|^2\exp(-G_{0,T})}{1-\exp(-G_{0,T})}\right] \leq \exp\left(\frac{R^2\exp(-G_{0,T})}{1-\exp(-G_{0,T})}\right)$$

Using convexity of χ^2-divergence then gives the result. For $G_{0,T} \geq \ln\left(\frac{4R^2}{\varepsilon^2}\right) \vee 1$, we have

$$\exp\left[\frac{R^2\exp(-G_{0,T})}{1-\exp(-G_{0,T})}\right] \leq \exp\left[\frac{\varepsilon^2/4}{1/2}\right] \leq \varepsilon^2. \qquad \square$$

Lemma B.15 (Subgaussian bound)**.** *Suppose $\widetilde{P}_0$ is supported on $B_R(0)$. Then for $X \sim \widetilde{p}_t$,*

$$\|X\|_{2,\psi_2} \leq \sqrt{\frac{e}{\ln 2}} \cdot \left(4m_tR + 6C_1\sigma_t\sqrt{d}\right) = O(\max\{R,\sqrt{d}\}),$$

where m_t, σ_t are as in (3) *and C_1 is an absolute constant.*

Proof. Let $Y \sim \widetilde{P}_0$ s.t. $X = m_tY + \sigma_t\xi$ for some $\xi \sim N(0, I_d)$ independent of Y. Define $U = \|X\|_2 := \left(\sum_{i=1}^d X_i^2\right)^{1/2}$, then for $p \geq 1$,

$$\begin{aligned}
\mathbb{E}|U|^p = \mathbb{E}\,\|X\|_2^p &\leq \mathbb{E}\left(\|m_tY\|_2 + \|\sigma_t\xi\|_2\right)^p \\
&\leq 2^{p-1}\mathbb{E}\left[\|m_tY\|_2^p + \|\sigma_t\xi\|_2^p\right] \\
&\leq 2^{p-1}\left[(m_tR)^p + \sigma_t^p \cdot 2^{p/2}\frac{\Gamma((d+p)/2)}{\Gamma(d/2)}\right] \\
&\leq 2^{p-1}\left[(m_tR)^p + C_1(\sqrt{2}\sigma_t)^p \cdot \left(d^{p/2} + p^{p/2}\right)\right]
\end{aligned}$$

where Γ is the commonly used gamma function and C_1 is an absolute constant. Therefore

$$(\mathbb{E}|U|^p)^{1/p} \leq 2m_tR + \sqrt{2}C_1\sigma_t(\sqrt{d} + \sqrt{p}) \leq K\sqrt{p},$$

where $K = 2m_tR + 3C_1\sigma_t\sqrt{d}$. Now consider $V = U/K$, then for some $\lambda > 0$ small enough, by Taylor expansion,

$$\mathbb{E}\left[e^{\lambda^2V^2}\right] = \mathbb{E}\left[1 + \sum_{p=1}^{\infty}\frac{\left(\lambda^2V^2\right)^p}{p!}\right] = 1 + \sum_{p=1}^{\infty}\frac{\lambda^{2p}\mathbb{E}\left[V^{2p}\right]}{p!}.$$

Note that $\mathbb{E}\left[V^{2p}\right] \leq (2p)^p$, while Stirling's approximation yields $p! \geq (p/e)^p$. Substituting these two bounds, we get

$$\mathbb{E}e^{\lambda^2V^2} \leq 1 + \sum_{p=1}^{\infty}\left(\frac{2\lambda^2p}{p/e}\right) = \sum_{p=0}^{\infty}(2e\lambda^2)^p = \frac{1}{1-2e\lambda^2},$$

provided that $2e\lambda^2 < 1$, in which case the geometric series above converges. To bound this quantity further, we can use the numeric inequality $1/(1-x) \leq e^{2x}$ which is valid for $x \in [0, 1/2]$. It follows that

$$\mathbb{E}e^{\lambda^2V^2} \leq e^{4e\lambda^2} \text{ for all } \lambda \text{ satisfying } |\lambda| \leq 1/2\sqrt{e}.$$

Now set $4e\lambda^2 = \ln 2$, then

$$\mathbb{E}\left[e^{\frac{\ln 2}{4eK^2}\|X\|_2^2}\right] \leq 2,$$

which implies that

$$\|X\|_{2,\psi_2} \leq \sqrt{\frac{4e}{\ln 2}}K = \sqrt{\frac{e}{\ln 2}} \cdot \left(4m_tR + 6C_1\sigma_t\sqrt{d}\right). \qquad \square$$

C Bounding the KL divergence

In this section, we bound the quantity $K = \mathrm{KL}(\psi_t q_t || p_t)$, where ψ_t is as in (8). While p_t is defined by the DDPM process, in this section we do *not* assume q_t is the density of the discretized process; rather, it is any density for which $\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right)$ and $\chi^2(q_t||p_t)$ are finite.

Lemma C.1. *Suppose that $\widetilde{P}_0$ is a probability measure on $\mathbb{R}^d$ such that*

$$\widetilde{P}_0 = \sum_{j=1}^m w_j \widetilde{P}_{j,0}, \tag{29}$$

where $w_j > 0$, $\sum_{j=1}^m w_j = 1$, and each $\widetilde{P}_{j,0}$ is a probability measure. For $t > 0$, let $\widetilde{p}_t$ and $\widetilde{p}_{j,t}$ be the densities obtained by running the forward DDPM process (1) *for time t, and $p_t = \widetilde{p}_{T-t}$, $p_{j,t} = \widetilde{p}_{j,T-t}$. Let $w_{\min} = \min_{1\le j\le m} w_j$ and suppose all the $\widetilde{P}_{j,t}$ satisfy a log-Sobolev constant with constant C_t. Then for any q_t, where ψ_t is as in* (8)

$$\mathrm{KL}(\psi_t q_t||p_t) \le \frac{2C_{T-t}}{\chi^2(q_t||p_t)+1} \cdot \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + \ln\left(\frac{1}{w_{\min}}\right).$$

While we need p_t to satisfy a log-Sobolev inequality to get a bound of the form $\frac{C}{\chi^2(q_t||p_t)+1}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right)$ ([LLT22, Lemma C.8]), we note that if we allow additive slack, it suffices for p_t to be a mixture of distributions satisfying a log-Sobolev inequality, with the *logarithm* of the minimum mixture weight bounded below. In Lemma C.2 we will see that we can almost decompose any distribution of bounded support in this manner, if we move a small amount of the mass.

Proof. Let $\overline{f}_t : [m] \to \mathbb{R}$ be the function

$$\overline{f}_t(j) = \int_{\mathbb{R}^d} \frac{\psi_t(x)q_t(x)}{p_t(x)} P_{j,t}(x)\,dx.$$

By decomposition of entropy and the fact that each $P_{i,t}$ satisfies LSI with constant C_{T-t},

$$\begin{aligned}
\mathrm{KL}(\psi_t q_t||p_t) &\le \mathrm{Ent}_{p_t}\left(\frac{\psi_t q_t}{p_t}\right) \\
&= \sum_{i=1}^m \int_{\mathbb{R}^d} w_i \,\mathrm{Ent}_{P_{i,t}}\left(\frac{\psi_t q_t}{p_t}\right) + \mathrm{Ent}_w(\overline{f}_t) \\
&\le \frac{C_t}{2}\sum_{i=1}^m w_i \mathscr{E}_{P_{i,t}}\left(\ln\frac{\psi_t q_t}{p_t}, \frac{\psi_t q_t}{p_t}\right) + \mathrm{Ent}_w(\overline{f}_t) \\
&\le \frac{C_t}{2}\mathscr{E}_{p_t}\left(\ln\frac{\psi_t q_t}{p_t}, \frac{\psi_t q_t}{p_t}\right) + \mathrm{Ent}_w(\overline{f}_t) \\
&= \frac{C_t}{2}\int_{\mathbb{R}^d}\left\|\nabla \ln\frac{\psi_t(x)q_t(x)}{p_t(x)}\right\|^2 \psi_t(x)q_t(x)\,dx + \mathrm{Ent}_w(\overline{f}_t) \\
&= 2C_t\int_{\mathbb{R}^d}\left\|\nabla \ln\frac{q_t(x)}{p_t(x)}\right\|^2 \psi_t(x)q_t(x)\,dx + \mathrm{Ent}_w(\overline{f}_t) \\
&= 2C_t\int_{\mathbb{R}^d}\left\|\nabla \frac{q_t(x)}{p_t(x)}\right\|^2 \frac{\psi_t(x)p_t(x)^2}{q_t(x)}\,dx + \mathrm{Ent}_w(\overline{f}_t) \\
&= \frac{2C_t}{\chi^2(q_t||p_t)+1}\cdot\int\left\|\nabla \frac{q_t(x)}{p_t(x)}\right\|^2 p_t(x)\,dx + \mathrm{Ent}_w(\overline{f}_t) \\
&\le \frac{2C_t}{\chi^2(q_t||p_t)+1}\cdot\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + \ln\left(\frac{1}{w_{\min}}\right),
\end{aligned}$$

where the last inequality follows from noting $w_j \overline{f}_t(j)$ is a probability mass function on $[m]$, so that $\overline{f}_t(j) \leq \frac{1}{w_j}$ and

$$\mathrm{Ent}_w(\overline{f}_t) = \sum_{j=1}^m w_j \overline{f}_t(j) \ln(\overline{f}_t(j)) \leq \sum_{j=1}^m w_j \overline{f}_t(j) \ln\left(\frac{1}{w_{\min}}\right) = \ln\left(\frac{1}{w_{\min}}\right). \qquad \square$$

Lemma C.2. *Suppose $0 < \varepsilon_K < \frac{1}{2}$, and that $\overline{P}_0$ is a probability measure such that $\overline{P}_0(\mathcal{M}) \geq 1 - \frac{\varepsilon_K}{8}$. Let $\mathcal{N}\left(\mathcal{M}, \frac{\sigma_t}{2}\right)$ denote the covering number of $\mathcal{M}$ with balls of radius σ_t. Given $\delta > 0$, there exists a distribution $\widetilde{P}_0$ such that $\chi^2(\widetilde{P}_0 || \overline{P}_0) \leq \varepsilon_K$ and considering the DDPM process started with $\widetilde{P}_0$, for all $0 \leq t \leq T - \delta$,*

$$\mathrm{KL}(\psi_t q_t || p_t) \leq \left(\frac{6(1+e)\sigma_{T-t}^2}{\chi^2(q_t||p_t)+1} \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + \ln\left(\frac{\mathcal{N}(\mathcal{M}, \sigma_\delta/2)}{\varepsilon_K}\right)\right).$$

In particular, for $\mathcal{M} = B_R(0)$ in $\mathbb{R}^d$,

$$\mathrm{KL}(\psi_t q_t || p_t) \leq \left(\frac{6(1+e)\sigma_{T-t}^2}{\chi^2(q_t||p_t)+1} \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + \ln\left(\frac{1}{\varepsilon_K}\right) + d \ln\left(1 + \frac{4R}{\sigma_\delta}\right)\right). \tag{30}$$

Proof. Partition $\mathcal{M}$ into disjoint subsets $\mathcal{M}_j$, $1 \leq j \leq N := \mathcal{N}(\mathcal{M}, \sigma_\delta/2)$ of diameter at most σ_δ, and decompose

$$\overline{P}_0 = w_* P_* + \sum_{j=1}^n \overline{w}_j \widetilde{P}_{j,0}$$

where p_j is supported on $\mathcal{M}_j$ and $P_* = \overline{P}_0(\cdot | \mathcal{M}^c)$. We will zero out the coefficients of all small components: let $Z = \sum_{j: w_j \geq \frac{\varepsilon_K}{8N}} w_j$ and

$$w_j = \begin{cases} \frac{\overline{w}_j}{Z}, & j \in [n], \overline{w}_j \geq \frac{\varepsilon_K}{8N} \\ 0, & \text{otherwise,} \end{cases}$$

and define

$$\widetilde{P}_0 = \sum_{j=1}^n w_j \widetilde{P}_{j,0}.$$

Note that $Z \geq 1 - \frac{\varepsilon_K}{8} - \sum_{j: w_j \leq \frac{\varepsilon_K}{8N}} \geq 1 - \frac{\varepsilon_K}{4}$. As probability distributions on $[m] \cup \{*\}$,

$$\chi^2(w || \overline{w}) \leq \left(\frac{1}{1 - \frac{\varepsilon_K}{4}}\right)^2 - 1 \leq \varepsilon_K,$$

and hence the same bound holds for $\chi^2(\widetilde{P}_0 || \overline{P}_0)$. Note each $M_{m_t \sharp} \widetilde{P}_{j,0}$ is supported on a set of diameter $m_t \sigma \leq \sigma$. By Theorem 1 of [CCN21], noting that

$$\chi^2(N(\mu_2, \Sigma) || N(\mu_1, \Sigma)) = \exp\left[(\mu_2 - \mu_1)^\top \Sigma^{-1} (\mu_2 - \mu_1)\right] \leq e$$

when $\Sigma = \sigma^2 I$ and $\|\mu_2 - \mu_1\| \leq \sigma$, $\widetilde{P}_{j,t} = (M_{m_t \sharp} \widetilde{P}_{j,0}) * \varphi_{\sigma^2}$ satisfies a log-Sobolev inequality with constant $6(1+e)\sigma_t^2$. The result then follows from Lemma C.1. For $\mathcal{M} = B_R(0)$, we use the bound $\mathcal{N}(B_R(0), \sigma_\delta/2) \leq \left(1 + \frac{4R}{\sigma_\delta}\right)^d$ [Ver18, Corollary 4.2.13]. $\square$

In the next section, we show that we can move a small amount of mass ε without significantly affecting the score function. This is necessary, as our guarantees on the score estimate are for the original distribution and not the perturbed one in Lemma C.2.

D The effect of perturbing the data distribution on the score function

In this section we consider the effect of perturbing the data distribution on the score function. The key observation is that the score function can be interpreted as the solution to an inference problem, that of recovering the original data point from a noisy sample, with data distribution as the given prior distribution. We show through a coupling argument that we can bound the difference between the score functions in terms of the distance between the two data distributions. This will allow us to "massage" the data distribution in order to optimally bound $\mathrm{KL}(\psi_t q_t || p_t)$ in Section C.

D.1 Perturbation under χ^2 error and truncation

We first give a general lemma on denoising error from a mismatched prior.

Lemma D.1 (Denoising error from mismatched prior)**.** *Let φ be a probability density on $\mathbb{R}^d$, and $P_{0,x}, P_{1,x}$ be measures on $\mathbb{R}^d$. For $i = 0, 1$, let P_i denote the joint distribution of $x_i \sim P_{i,x}$ and $y_i = x_i + \xi_i$ where $\xi_i \sim \varphi$, and let $P_{i,y}$ denote the marginal distribution of y. Let*

$$m^{(k)}(\varepsilon) := \sup_{0 \le f \le 1, \int_{\mathbb{R}^d} f\varphi\, dx \le \varepsilon} \int_{\mathbb{R}^d} f(x) \|x\|^k \varphi(x)\, dx.$$

Let $\varepsilon_{\mathrm{TV}} = \mathrm{TV}(P_{0,x}, P_{1,x})$ and $\varepsilon_\chi^2 = \chi^2(P_{0,x} || P_{1,x})$. Then

$$\int_{\mathbb{R}^d} P_{0,y}(dy_0) \left\| \int_{\mathbb{R}^d} x_0 P_0(dx_0|y_0) - \int_{\mathbb{R}^d} x_1 P_1(dx_1|y_0) \right\|^2$$
$$\le 8m^{(2)}(\varepsilon_{\mathrm{TV}}) + \varepsilon_\chi \sqrt{m^{(4)}(\varepsilon_{\mathrm{TV}})}$$

For $\varphi = \varphi_{\sigma^2}$, the upper bound is $O\left(\sigma^2 \varepsilon_\chi \left(d + \ln\left(\frac{1}{\varepsilon_{\mathrm{TV}}}\right)\right)\right)$.

Note the tricky part of the proof is to deal with $P_1(dx_1|y_0)$, which can be thought of as inferring x assuming the incorrect prior $P_{1,x}$, rather than the actual prior $P_{0,x}$.

Proof. For notational clarity, we will denote draws from the conditional distribution as $\widehat{x}_0$ and $\widehat{x}_1$, for example $P_0(d\widehat{x}_0|y_0)$. Let $r_i(y) = \int_{\mathbb{R}^d} (\widehat{x}_i - y) P_i(d\widehat{x}_i|y)$. Let $P_{0,1}$ be a coupling of $(x_0, y_0 = x_0 + \xi_0, x_1, y_1 = y_1 + \xi_1)$ such that $x_0 = x_1$ with probability $1 - \varepsilon_{\mathrm{TV}}$ and $\xi_0 = \xi_1$ with probability 1. We have

$$\int_{\mathbb{R}^d} P_{0,y}(dy_0) \|r_0(y_0) - r_1(y_0)\|^2 = \underbrace{\int_{\{y_0 = y_1\}} P_{0,1,y}(dy_0, dy_1) \|r_0(y_0) - r_1(y_0)\|^2}_{(I)}$$
$$+ \underbrace{\int_{\{y_0 \neq y_1\}} P_{0,1,y}(dy_0, dy_1) \|r_0(y_0) - r_1(y_0)\|^2}_{(II)}.$$

Define a measure Q (not necessarily a probability measure) on $\mathbb{R}^d$ by

$$Q(A) := P_{0,1}(y_0 \in A \text{ and } y_0 = y_1).$$

Note that

$$Q(A) \le \min\{P_{0,y}(A), P_{1,y}(A)\},$$

so Q is absolutely continuous with respect to $P_{0,y}$ and $P_{1,y}$, and by assumption on the coupling,

$$Q(\mathbb{R}^d) \ge 1 - \varepsilon_{\mathrm{TV}}. \tag{31}$$

Under $P_{0,1}$, when $y_0 = y_1$, we can couple $P_0(d\widehat{x}_0|y_0)$ and $P_1(d\widehat{x}_1|y_0)$ so that $x_0 = x_1$ with probability $\min\left\{\frac{dQ}{dP_{0,y}}, \frac{dQ}{dP_{1,y}}\right\}$. Let $\widehat{P}(d\widehat{x}_0, d\widehat{x}_1|y_0)$ denote this coupled distribution. Then as in

Lemma D.5,

$$(I) \le \int_{\{y_0=y_1\}} P_{0,1,y}(dy_0, dy_1) \left\| \int_{\{\widehat{x}_0 \neq \widehat{x}_1\}} ((\widehat{x}_0 - y_0) - (\widehat{x}_1 - y_0)) \widehat{P}(d\widehat{x}_0, d\widehat{x}_1 | y_0) \right\|^2$$
$$\le 2 \int_{\mathbb{R}^d} P_{0,1,y}(dy_0, dy_1) \left(\int_{\{\widehat{x}_0 \neq \widehat{x}_1\}} \|\xi_0\|^2 \, \widehat{P}(d\widehat{x}_0, d\widehat{x}_1 | y_0) + \int_{\{\widehat{x}_0 \neq \widehat{x}_1\}} \|\xi_1\|^2 \, \widehat{P}(d\widehat{x}_0, d\widehat{x}_1 | y_1) \right)$$

We bound this by first bounding

$$\int_{\mathbb{R}^d} P_{0,1,y}(dy_1, dy_2) \widehat{P}(\widehat{x}_0 \neq \widehat{x}_1) \le \int_{\mathbb{R}^d} P_{0,y}(dy) \max\left\{1 - \frac{dQ}{dP_{0,y}}, 1 - \frac{dQ}{dP_{1,y}}\right\} \le 2\varepsilon_{\mathrm{TV}}, \quad (32)$$

which follows from the two inequalities (using (31))

$$\int_{\mathbb{R}^d} P_{0,y}(dy) \left(1 - \frac{dQ}{dP_{0,y}}\right) = 1 - Q(\mathbb{R}^d) \le \varepsilon_{\mathrm{TV}}$$
$$\int_{\mathbb{R}^d} P_{0,y}(dy) \left(1 - \frac{dQ}{dP_{1,y}}\right) \le \int_{\mathbb{R}^d} P_{1,y}(dy) \left(1 - \frac{dQ}{dP_{1,y}}\right) + \mathrm{TV}(P_{0,y}, P_{1,y})$$
$$\le (1 - Q(\mathbb{R}^d)) + \varepsilon_{\mathrm{TV}} \le 2\varepsilon_{\mathrm{TV}}.$$

From (32), and the fact that the distribution of (x_i, y_i) is the same as $(\widehat{x}_i, y_i)$ by Nishimori's identity, we obtain

$$(I) \le 2(m^{(2)}(2\varepsilon_{\mathrm{TV}}) + m^{(2)}(2\varepsilon_{\mathrm{TV}})) = 4m^{(2)}(\varepsilon_{\mathrm{TV}}).$$

Now for the second term (II),

$$(II) \le 2 \int_{\{y_0 \neq y_1\}} P_{0,1,y}(dy_0, dy_1)(\|r_0(y_0)\|^2 + \|r_1(y_0)\|^2).$$

The first term satisfies $\int_{\{y_0 \neq y_1\}} P_{0,1,y}(dy_0, dy_1) \|r_0(y_0)\|^2 \le m^{(2)}(\varepsilon_{\mathrm{TV}})$. For the second term, we note that Cauchy-Schwarz gives for any measures P and Q that

$$\int_\Omega f(x)P(dx) \le \int_\Omega f(x)Q(dx) + \int_\Omega \left(\frac{dP}{dQ} - 1\right) f(x)Q(dx)$$
$$\le \int_\Omega f(x)Q(dx) + \sqrt{\chi^2(P||Q) \int_\Omega f(x)^2 Q(dx)}$$

to switch from the measure $P_{0,y}$ to $P_{1,y}$:

$$\int_{\{y_0 \neq y_1\}} P_{0,1,y}(dy_0) \|r_1(y_0)\|^2 = \int_{\mathbb{R}^n} P_{0,y}(dy_0) P_{0,1,y}(y_0 \neq y_1 | y_0) \|r_1(y_0)\|^2$$
$$\le \int_{\mathbb{R}^n} P_{1,y}(dy_0) P_{0,1,y}(y_0 \neq y_1 | y_0) \|r_1(y_0)\|^2 + \sqrt{\chi^2(P_{0,y}||P_{1,y}) \int P_{1,y}(dy_0) P_{0,1,y}(y_0 \neq y_1 | y_0) \|r_1(y_0)\|^4}$$

(Note that intentionally, the measure is $P_{1,y}$, though we use y_0 for the variable.) Hence,

$$\int_{\mathbb{R}^n} P_{1,y}(dy_0) P_{0,1,y}(y_0 \neq y_1 | y_0) \le \mathrm{TV}(P_{0,y}, P_{1,y}) + \int_{\mathbb{R}^n} P_{0,y}(dy_0) P_{0,1,y}(y_0 \neq y_1 | y_0) \le 2\varepsilon_{\mathrm{TV}}$$

so

$$\int_{\{y_0 \neq y_1\}} P_{0,1,y}(dy_0) \|r_1(y_0)\|^2 \le m^{(2)}(2\varepsilon_{\mathrm{TV}}) + \sqrt{\chi^2(P_{0,x}||P_{1,x}) m^{(4)}(2\varepsilon_{\mathrm{TV}})},$$

where we used the data processing inequality.

For $\varphi = \varphi_{\sigma^2}$, we obtain by Lemma D.6 that the bound is

$$O\left(\sigma^2(\varepsilon_{\mathrm{TV}} + \varepsilon_\chi \varepsilon_{\mathrm{TV}}^{1/2}) \left(d + \ln\left(\frac{1}{\varepsilon_{\mathrm{TV}}}\right)\right)\right) = O\left(\sigma^2 \varepsilon_\chi \left(d + \ln\left(\frac{1}{\varepsilon_{\mathrm{TV}}}\right)\right)\right). \qquad \square$$

We use this lemma to obtain a bound on the L^2 score error under perturbation of the distribution, by interpreting the score as the solution to a de-noising problem.

Lemma D.2 (L^2 score error under perturbation)**.** *Let $\widetilde{P}^{(0)} = \widetilde{P}_0^{(0)}$ and $\widetilde{P}^{(1)} = \widetilde{P}_0^{(1)}$ be two probability distributions on $\mathbb{R}^d$ such that $\chi^2(\widetilde{P}^{(1)}||\widetilde{P}^{(0)}) \le \varepsilon_\chi^2 \le 1$.*

1. *For any $\sigma > 0$,*

$$\int \left\|\nabla \ln(\widetilde{P}^{(0)} * \varphi_{\sigma^2})(x) - \nabla \ln(\widetilde{P}^{(1)} * \varphi_{\sigma^2})(x)\right\|^2 (\widetilde{P}^{(1)} * \varphi_{\sigma^2})(dx) = O\left(\frac{\varepsilon_\chi\left(d + \ln\left(\frac{1}{\varepsilon_\chi}\right)\right)}{\sigma^2}\right).$$

2. *Let $\widetilde{p}_t^{(i)}$ be the density resulting from running* (1) *starting from $\widetilde{P}^{(i)}$, and let σ_t be as in* (3)*. Then for any $t > 0$,*

$$\int \left\|\nabla \ln \widetilde{p}_t^{(0)}(x) - \nabla \ln \widetilde{p}_t^{(1)}(x)\right\|^2 \widetilde{p}_t^{(1)}(x)\, dx = O\left(\frac{\varepsilon_\chi\left(d + \ln\left(\frac{1}{\varepsilon_\chi}\right)\right)}{\sigma_t^2}\right).$$

Proof. For part 1, note by (27) that

$$\nabla \ln(\widetilde{P}^{(i)} * \varphi_{\sigma^2})(y) = \frac{1}{\sigma^2}\mathbb{E}_{\widetilde{P}^{(i)}_{y,\sigma^2}}(x - y),$$

where $\widetilde{P}^{(i)}_{y,\sigma^2}$ is the "tilted" probability distribution defined by

$$\frac{d\widetilde{P}^{(i)}_{y,\sigma^2}}{d\widetilde{P}^{(i)}}(x) \propto e^{-\frac{\|x-y\|^2}{2\sigma^2}}.$$

By Bayes's rule, this can be viewed as the conditional probability that $x_0 = x$ given $x_t = y$, where $x_0 \sim \widetilde{P}^{(i)}$ and $y = x_0 + \sigma\xi$, $\xi \sim N(0, I_d)$. Hence this fits in the framework of Lemma D.1 and

$$\begin{aligned}
&\int \left\|\nabla \ln(\widetilde{P}^{(0)} * \varphi_{\sigma^2})(y) - \nabla \ln(\widetilde{P}^{(1)} * \varphi_{\sigma^2})(y)\right\|^2 (\widetilde{P}^{(1)} * \varphi_{\sigma^2})(dy) \\
&= \frac{1}{\sigma^4}\int_{\mathbb{R}^d} \left\|\mathbb{E}_{\widetilde{P}^{(0)}_{y,t}}[x] - \mathbb{E}_{\widetilde{P}^{(1)}_{y,t}}[x]\right\|^2 (\widetilde{P}^{(1)} * \varphi_{\sigma^2})(dy) \\
&= O\left(\frac{1}{\sigma^4}\sigma^2\varepsilon_\chi\left(d + \ln\left(\frac{1}{\varepsilon_{\mathrm{TV}}}\right)\right)\right),
\end{aligned}$$

giving the result.

For part 2, note that $\widetilde{p}_t^{(i)} = (M_{m_t\sharp}\widetilde{P}^{(i)}) * \varphi_{\sigma_t^2}$. Applying part 1 with $\widetilde{P}^{(i)} \hookleftarrow M_{m_t\sharp}\widetilde{P}^{(i)}$ (which preserves χ^2-divergence) and $\sigma = \sigma_t$ gives the result.

□

Finally, we argue that a score estimate that is accurate with respect to $\widetilde{p}_t^{(1)}$ will still be accurate with respect to $\widetilde{p}_t^{(0)}$, with high probability. When using this lemma, we will substitute in the bound from Lemma D.2.

Lemma D.3. *Let $\widetilde{P}_0^{(0)}$ and $\widetilde{P}_0^{(1)}$ be two probability distributions on $\mathbb{R}^d$ with TV distance ε. Suppose the estimated score function $s_t(x)$ satisfies*

$$\left\|\nabla \ln \widetilde{p}_t^{(0)} - s_t\right\|^2_{L^2(\widetilde{p}_t^{(0)})} = \mathbb{E}_{\widetilde{p}_t^{(0)}}\left[\left\|\nabla \ln \widetilde{p}_t^{(0)}(x) - s_t(x)\right\|^2\right] \le \varepsilon_t^2$$

for $t \in (0, T]$, and $\nabla \ln \widetilde{p}_t^{(0)}$ is L_t-Lipschitz. Then for $t \in (0, T]$ and any $\varepsilon_\infty > 0$,

$$P_{\widetilde{p}_t^{(1)}}\left(\left\|s_t - \nabla \ln \widetilde{p}_t^{(1)}\right\| \ge \varepsilon_\infty\right) \le \varepsilon + \frac{4}{\varepsilon_\infty^2}\cdot\left[\varepsilon_t^2 + \int \left\|\nabla \ln \widetilde{p}_t^{(1)}(x) - \nabla \ln \widetilde{p}_t^{(0)}(x)\right\|^2 \widetilde{p}_t^{(1)}(x)\, dx\right].$$

Proof. We have

$$
\begin{aligned}
&P_{\widetilde{p}_t^{(1)}}\left(\left\|s_t - \nabla \ln \widetilde{p}_t^{(1)}\right\| \geq \varepsilon_\infty\right) \\
&\leq P_{\widetilde{p}_t^{(1)}}\left(\left\|s_t - \nabla \ln \widetilde{p}_t^{(0)}\right\| \geq \varepsilon_\infty/2\right) + P_{\widetilde{p}_t^{(1)}}\left(\left\|\nabla \ln \widetilde{p}_t^{(0)} - \nabla \ln \widetilde{p}_t^{(1)}\right\| \geq \varepsilon_\infty/2\right) \\
&\leq \mathrm{TV}(\widetilde{p}_t^{(0)}, \widetilde{p}_t^{(1)}) + P_{\widetilde{p}_t^{(0)}}\left(\left\|s_t - \nabla \ln \widetilde{p}_t^{(0)}\right\| \geq \varepsilon_\infty/2\right) + P_{\widetilde{p}_t^{(1)}}\left(\left\|\nabla \ln \widetilde{p}_t^{(0)} - \nabla \ln \widetilde{p}_t^{(1)}\right\| \geq \varepsilon_\infty/2\right).
\end{aligned}
$$

The first term is bounded by $\mathrm{TV}(\widetilde{P}^{(0)}, \widetilde{P}^{(1)}) \leq \varepsilon$. For the second term, by Chebyshev's Inequality,

$$
P_{\widetilde{p}_t^{(0)}}\left(\left\|s_t - \nabla \ln \widetilde{p}_t^{(0)}\right\| \geq \varepsilon_1/2\right) \leq \frac{4}{\varepsilon_\infty^2}\mathbb{E}_{\widetilde{p}_t^{(0)}}\left[\left\|s_t - \nabla \ln \widetilde{p}_t^{(0)}\right\|^2\right] \leq \frac{4\varepsilon_t^2}{\varepsilon_\infty^2};
$$

For the last term, again by Chebyshev's Inequality,

$$
P_{\widetilde{p}_t^{(1)}}\left(\left\|\nabla \ln \widetilde{p}_t^{(0)} - \nabla \ln \widetilde{p}_t^{(1)}\right\| \geq \varepsilon_\infty/2\right) \leq \frac{4}{\varepsilon_1^2}\int \left\|\nabla \ln \widetilde{p}_t^{(1)}(x) - \nabla \ln \widetilde{p}_t^{(0)}(x)\right\|^2 \widetilde{p}_t^{(1)}(x)dx.
$$

We conclude the proof by combining the these three inequalities. □

Finally, we will need the following to obtain a TV error bound to $\widetilde{p}_0$ in Theorem 3.3.

Lemma D.4. *Suppose that $\widetilde{p}_0 \propto e^{-V(x)}$ is a probability density on $\mathbb{R}^d$ with bounded first moment $\mathbb{E}_{\widetilde{p}_0}\|X\|$, and V is L-smooth. Then for $t > 0$ such that $\alpha_t\sigma_t \leq \frac{1}{2L}$, we have*

$$
\mathrm{TV}(\widetilde{p}_t, \widetilde{p}_0) \leq 2\left(\alpha_t - 1\right) \cdot \left(L\mathbb{E}_{\widetilde{p}_0}\|X\| + d\right) + \frac{3}{2}dL\alpha_t\sigma_t.
$$

Here $\alpha_t = 1/m_t$ and σ_t are defined in (3)*. In particular, $TV(\widetilde{p}_\delta, \widetilde{p}_0) \leq \varepsilon_{\mathrm{TV}}$ if $\delta = O(\frac{\varepsilon_{TV}^2}{R^2L^2})$ and $R = \max\left\{\sqrt{d}, \mathbb{E}_{\widetilde{p}_0}\|X\|\right\}$.*

Proof. Without loss of generality, we assume that $\widetilde{p}_0(x) = e^{-V(x)}$. Note that $\widetilde{p}_t(x) = \int \alpha_t^d \widetilde{p}_0(\alpha_t y)\varphi_{\sigma_t^2}(x-y)\,dy$. Let $\widetilde{q}_t(x) := \alpha_t^d\widetilde{p}_0(\alpha_t x)$, which is also a probability density on $\mathbb{R}^d$. Then by the triangle inequality,

$$
\mathrm{TV}(\widetilde{p}_t, \widetilde{p}_0) \leq \mathrm{TV}(\widetilde{p}_t, \widetilde{q}_t) + \mathrm{TV}(\widetilde{q}_t, \widetilde{p}_0).
$$

For the second term,

$$
\begin{aligned}
|\widetilde{q}_t(x) - \widetilde{p}_0(x)| &= \left|\alpha_t^d\widetilde{p}_0(\alpha_t x) - \widetilde{p}_0(x)\right| \\
&= \left|e^{-V(\alpha_t x) + d\ln\alpha_t} - e^{-V(x)}\right| \\
&\leq \max\left\{e^{-V(x)}, e^{-V(\alpha_t x)+d\ln\alpha_t}\right\} \cdot \left(1 - e^{-|V(x)-V(\alpha_t x)+d\ln\alpha_t|}\right) \\
&\leq \left(\widetilde{p}_0(x) + \widetilde{q}_t(x)\right) \cdot \left(|V(x) - V(\alpha_t x)| + d\ln\alpha_t\right) \\
&\leq \left(\widetilde{p}_0(x) + \widetilde{q}_t(x)\right) \cdot \left[L\|x\|\left(\alpha_t - 1\right) + d\ln\alpha_t\right],
\end{aligned}
$$

where in the second inequality, we use the fact that $1 - e^x \leq |x|$ for all $x \leq 0$. Thus

$$
\begin{aligned}
\mathrm{TV}(\widetilde{q}_t(x), \widetilde{p}_0(x)) &= \frac{1}{2}\int |\widetilde{q}_t(x) - \widetilde{p}_0(x)|\,dx \\
&\leq \int \left[L\left(\alpha_t - 1\right)\|x\| + d\ln\alpha_t\right]\widetilde{p}_0(x)\,dx + \int \left[L\left(\alpha_t - 1\right)\|x\| + d\ln\alpha_t\right]\widetilde{q}_t(x)\,dx \\
&\leq L(\alpha_t - 1)\left(\int \|x\|\,\widetilde{p}_0(x)dx + \int \|x\|\,\widetilde{q}_t(x)dx\right) + 2d\ln\alpha_t \\
&\leq 2L(\alpha_t - 1)\int \|x\|\,\widetilde{p}_0(x)dx + 2d\ln\alpha_t.
\end{aligned}
$$

Now for the first term,

$$
\widetilde{p}_t(x) - \widetilde{q}_t(x) = \int \widetilde{q}_t(x-y)\varphi_{\sigma_t^2}(y)\,dy - \widetilde{q}_t(x) = \int \left(\widetilde{q}_t(x - \sigma_t y) - \widetilde{q}_t(x)\right)\varphi(y)dy,
$$

where $\varphi(y)$ is the density of the d-dimensional standard Gaussian distribution. Apply Minkowski's inequality for integrals:

$$\begin{aligned}
\int |\widetilde{p}_t(x) - \widetilde{q}_t(x)|\, dx &= \int \left| \int \left(\widetilde{q}_t(x - \sigma_t y) - \widetilde{q}_t(x)\right) \varphi(y) dy \right| dx \\
&\leq \int \left[\int |\widetilde{q}_t(x - \sigma_t y) - \widetilde{q}_t(x)|\, dx \right] \varphi(y)\, dy \\
&\leq \int \left[\int \left(e^{L\|\alpha_t \sigma_t y\|} - 1 \right) \widetilde{q}_t(x) dx \right] \varphi(y)\, dy \\
&= \int \left(e^{L\|\alpha_t \sigma_t y\|} - 1 \right) \varphi(y)\, dy \\
&= (2\pi)^{-d/2} \int e^{L\alpha_t \sigma_t \|y\| - \frac{\|y\|^2}{2}}\, dy - 1 \\
&\leq (2\pi)^{-d/2} \int e^{\left[-\frac{1}{2} + (L\alpha_t\sigma_t)^2\right]\|y\|^2} dy + L\alpha_t\sigma_t \int \|y\|\, \varphi(y)\, dy - 1 \\
&\leq \left[\frac{1}{1 - 2\left(L\alpha_t\sigma_t\right)^2} \right]^{d/2} + \sqrt{d} L\alpha_t\sigma_t - 1 \\
&\leq e^{2d(L\alpha_t\sigma_t)^2} - 1 + \sqrt{d} L\alpha_t\sigma_t \\
&\leq 4d\left(L\alpha_t\sigma_t\right)^2 + \sqrt{d} L\alpha_t\sigma_t,
\end{aligned}$$

where in the third inequality, we use the elementary inequality $e^x \leq x + e^{x^2}$, which is valid for all $x \in \mathbb{R}$, and in the fifth inequality, we use $\frac{1}{1-2x} \leq e^{4x}$, which holds for $x \in [0, 1/3]$. Hence if $L\alpha_t\sigma_t \leq 1/2$, we have

$$\mathrm{TV}(\widetilde{p}_t, \widetilde{q}_t) \leq \frac{1}{2} \int |\widetilde{p}_t(x) - \widetilde{q}_t(x)|\, dx \leq \frac{3}{2} dL\alpha_t\sigma_t.$$

Now we conclude the proof by combining the bounds for $\mathrm{TV}(\widetilde{p}_t, \widetilde{q}_t)$ and $\mathrm{TV}(\widetilde{p}_0, \widetilde{q}_t)$:

$$\begin{aligned}
\mathrm{TV}(\widetilde{p}_t, \widetilde{p}_0) &\leq \mathrm{TV}(\widetilde{p}_t, \widetilde{q}_t) + \mathrm{TV}(\widetilde{q}_t, \widetilde{p}_0) \\
&\leq 2L(\alpha_t - 1) \int \|x\|\, \widetilde{p}_0(x) dx + 2d \ln \alpha_t + \frac{3}{2} dL\alpha_t\sigma_t \\
&\leq 2\left(\alpha_t - 1\right) \cdot \left(L\mathbb{E}_{\widetilde{p}_0} \|X\| + d\right) + \frac{3}{2} dL\alpha_t\sigma_t,
\end{aligned}$$

where we use the fact that $\ln x \leq x - 1$ for all $x \geq 1$. Recall that $\alpha_t = 1/m_t = e^{t/2}$ and $\sigma_t^2 = 1 - e^{-t}$ when $g \equiv 1$. It suffices for

$$\max\left\{ 2\left(L\mathbb{E}_{\widetilde{p}_0} \|X\| + d\right)\left(\alpha_\delta - 1\right), \frac{3}{2} dL\alpha_\delta\sigma_\delta \right\} \leq \frac{\varepsilon_{\mathrm{TV}}}{2},$$

which is implied by

$$\delta \precsim \min\left\{ \frac{\varepsilon_{\mathrm{TV}}}{L\mathbb{E}_{\widetilde{p}_0} \|X\| + d}, \frac{\varepsilon_{\mathrm{TV}}^2}{d^2 L^2} \right\} \asymp \frac{\varepsilon_{\mathrm{TV}}^2}{R^2 L^2},$$

for appropriate constants, as $R \geq \max\left\{ \sqrt{d}, \mathbb{E}_{\widetilde{p}_0} \|X\| \right\}$. □

D.2 Perturbation under TV error

Although we will not need it in our proof, we note that we can derive a similar perturbation result under TV error, which might be of independent interest.

Lemma D.5. *Let $K(x, dy)$ be a probability kernel on $\mathbb{R}^d$, let $P_{0,x}, P_{1,x}$ be measures on $\mathbb{R}^d$. Let P_i denote the joint distribution of $x_i \sim P_{i,x}$ and $y_i \sim K(x_i, \cdot)$, and let $P_{i,y}$ denote the marginal distribution of y. Suppose there is a coupling $P_{0,1}$ of $(x_0, y_0) \sim P_0$ and $(x_1, y_1) \sim P_1$ such that*

- $x_0 = x_1$ *with probability* $1-\varepsilon$,
- $x_0 = x_1$ *implies* $y_0 = y_1$, *and*
- $\mathbb{E}[\|y_0 - y_1\|^2] \le \varepsilon_{\mathrm{W}}^2$.

Define the tail error by

$$m_i(\varepsilon) := \sup_{0 \le f \le 1, \int_{\mathbb{R}^d} f\varphi\, dx \le \varepsilon} \int_{\mathbb{R}^d} f(x)\, \|x\|^2\, P_i(\, dx).$$

Let $r_i(y) = \int_{\mathbb{R}^d} x_i P_i(dx_i|y)$, *and suppose that* $r_1(y) = \int_{\mathbb{R}^d} x_1 P_1(dx_1|y)$ *is* L_1*-Lipschitz. Then*

$$\begin{aligned}
&\int_{\mathbb{R}^d} P_{0,y}(dy_0) \left\| \int_{\mathbb{R}^d} x_0 P_0(dx_0|y_0) - \int_{\mathbb{R}^d} x_1 P_1(dx_1|y_0) \right\|^2 \\
&\quad \le 4(m_0(2\varepsilon) + m_0(\varepsilon) + m_1(2\varepsilon) + m_1(\varepsilon)) + 2L_1^2 \varepsilon_{\mathrm{W}}^2 \\
&\quad \le 4(m_0(2\varepsilon) + m_1(2\varepsilon)) + 4(1 + L_1^2)(m_0(\varepsilon) + m_1(\varepsilon)).
\end{aligned}$$

Proof. For notational clarity, we will denote draws from the conditional distribution as $\widehat{x}_0$ and $\widehat{x}_1$, for example $P_0(d\widehat{x}_0|y_0)$. We have

$$\begin{aligned}
\int_{\mathbb{R}^d} P_{0,y}(dy_0)\, \|r_0(y_0) - r_1(y_0)\|^2 \le{}& 2 \underbrace{\int_{\mathbb{R}^d \times \mathbb{R}^d} P_{0,1,y}(dy_0, dy_1)\, \|r_0(y_0) - r_1(y_1)\|^2}_{(I)} \\
&+ 2 \underbrace{\int_{\mathbb{R}^d \times \mathbb{R}^d} P_{0,1,y}(dy_0, dy_1)\, \|r_1(y_1) - r_1(y_0)\|^2}_{(II)}.
\end{aligned}$$

For the first term (I), we split it as

$$(I) \le \underbrace{\int_{\{y_0 = y_1\}} P_{0,1,y}(dy_0, dy_1)\, \|r_0(y_0) - r_1(y_0)\|^2}_{(i)} + \underbrace{\int_{\{y_0 \ne y_1\}} P_{0,1,y}(dy_0, dy_1)\, \|r_0(y_0) - r_1(y_1)\|^2}_{(ii)}.$$

Define the measure Q on $\mathbb{R}^d$ by

$$Q(A) := P_{0,1}(y_0 \in A \text{ and } y_0 = y_1).$$

As in Lemma D.2, under $P_{0,1}$, when $y_0 = y_1$, we can couple $P_0(d\widehat{x}_0|y_0)$ and $P_1(d\widehat{x}_1|y_0)$ so that $x_0 = x_1$ with probability $\min\left\{\frac{dQ}{dP_{0,y}}, \frac{dQ}{dP_{1,y}}\right\}$. Let $\widehat{P}(d\widehat{x}_0, d\widehat{x}_1|y_0)$ denote this coupled distribution. Then

$$\begin{aligned}
(i) &\le \int_{\{y_0 = y_1\}} P_{0,1,y}(dy_0, dy_1) \left\| \int_{\{\widehat{x}_0 \ne \widehat{x}_1\}} (\widehat{x}_0 - \widehat{x}_1) \widehat{P}(d\widehat{x}_0, d\widehat{x}_1|y_0) \right\|^2 \\
&\le 2 \int_{\mathbb{R}^d} P_{0,1,y}(dy_0, dy_1) \left(\int_{\{\widehat{x}_0 \ne \widehat{x}_1\}} \|\widehat{x}_0\|^2\, \widehat{P}(d\widehat{x}_0, d\widehat{x}_1|y_0) + \int_{\{\widehat{x}_0 \ne \widehat{x}_1\}} \|\widehat{x}_1\|^2\, \widehat{P}(d\widehat{x}_0, d\widehat{x}_1|y_1) \right) \\
&\le 2(m_0(2\varepsilon) + m_1(2\varepsilon))
\end{aligned}$$

as in Lemma D.2. Now

$$(ii) \le 2 \int_{\{y_0 \ne y_1\}} P_{0,1,y}(dy_0, dy_1)(\|r_0(y_0)\|^2 + \|r_1(y_1)\|^2) \le 2(m_0(\varepsilon) + m_1(\varepsilon)).$$

Finally, for the second term (II), we use the fact that r_1 is L_1 Lipschitz and the coupling to conclude

$$(II) \le \int_{\mathbb{R}^d} P_{0,1,y}(dy_0, dy_1) L_1^2\, \|y_0 - y_1\|^2 \le L_1^2 \varepsilon_{\mathrm{W}}^2.$$

We conclude the proof by combining the inequalities for (i), (ii), and (II).

For the second upper bound, we note that

$$\mathbb{E}[\|y_0 - y_1\|^2] \le 2(\mathbb{E}[\|y_0\|^2] + \mathbb{E}[\|y_1\|^2]) \le 2(m_0(\varepsilon) + m_1(\varepsilon)). \qquad \square$$

D.3 Gaussian tail calculation

We use the following Gaussian tail calculation in the proof of Lemma D.2.

Lemma D.6. *Let μ be the standard Gaussian measure on $N(0, I_d)$. Then*

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

$$\sup_{\mu(A)\le\varepsilon}\int_A \|x\|^4\,\mu(dx) \le \varepsilon\left(2d+3\ln\left(\frac{1}{\varepsilon}\right)\right)^2 + 3\varepsilon\left(2d+3\ln\left(\frac{1}{\varepsilon}\right)\right)+9\varepsilon = O\left(\varepsilon\left(d^2+\ln\left(\frac{1}{\varepsilon}\right)^2\right)\right).$$

Proof. By the χ^2 tail bound in [LM00], for $t \ge 0$,

$$\mu(\|X\|^2 \ge 2d+3t) \le \mathbb{P}(\|X\|^2 \ge d+2\sqrt{dt}+2t) \le e^{-t}, \tag{33}$$

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

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

and

$$\begin{aligned}\sup_{\mu(A)\le\varepsilon}\int_A \|x\|^4\,\mu(dx) &\le \sup_{P_Y(A)\le\varepsilon}\int_A y^2P_Y(dy) = \int_{2d+3\ln\left(\frac{1}{\varepsilon}\right)}^{\infty} y^2\,dF(y)\\ &= \varepsilon\left(2d+3\ln\left(\frac{1}{\varepsilon}\right)\right)^2 + \int_{2d+3\ln\left(\frac{1}{\varepsilon}\right)}^{\infty} 2ye^{-\frac{y-2d}{3}}\,dy\\ &= \varepsilon\left(2d+3\ln\left(\frac{1}{\varepsilon}\right)\right)^2 - \left[3ye^{-\frac{y-2d}{3}}\right]\Big|_{2d+3\ln\left(\frac{1}{\varepsilon}\right)}^{\infty} + \int_{2d+3\ln\left(\frac{1}{\varepsilon}\right)}^{\infty} 3e^{-\frac{y-2d}{3}}\,dy\\ &= \varepsilon\left(2d+3\ln\left(\frac{1}{\varepsilon}\right)\right)^2 + 3\varepsilon\left(2d+3\ln\left(\frac{1}{\varepsilon}\right)\right)+9\varepsilon.\end{aligned}$$

□

E Guarantees under L^2-accurate score estimate

We will state our results under a more general tail bound assumption.

Assumption 6 (Tail bound)**.** *$R : [0,1] \to [0,\infty)$ is a function such that $P_{\text{data}}(B_{R(\varepsilon)}(0)) \ge 1-\varepsilon$.*

Our result will require $R(\varepsilon)$ to grow at most as a sufficiently small power of ε^{-1} as $\varepsilon \to 0$; in particular, this holds for subexponential distributions. By taking R to be a constant function, this contains the assumption of bounded support (Assumption 2) as a special case.

E.1 TV error guarantees

We follow the framework of [LLT22] to convert guarantees under L^∞-accurate score estimate, to guarantees under L^2-accurate score estimate.

Theorem E.1 ([LLT22, Theorem 4.1])**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\{\mathcal{F}_n\}$ be a filtration of the sigma field $\mathcal{F}$. Suppose $X_n \sim p_n$, $Z_n \sim q_n$, and $\overline{Z}_n \sim \overline{q}_n$ are $\mathcal{F}_n$-adapted random processes taking values in Ω, and $B_n \subseteq \Omega$ are sets such that the following hold for every $n \in \mathbb{N}_0$.*

1. *If $Z_k \in B_k^c$ for all $0 \le k \le n-1$, then $Z_n = \overline{Z}_n$.*
2. *$\chi^2(\overline{q}_n || p_n) \le D_n^2$.*

3. $\mathbb{P}(X_n \in B_n) \le \delta_n$.

Then the following hold.

$$\mathrm{TV}(q_n, \overline{q}_n) \le \sum_{k=0}^{n-1} (D_k^2+1)^{1/2} \delta_k^{1/2} \qquad \mathrm{TV}(p_n, q_n) \le D_n + \sum_{k=0}^{n-1} (D_k^2+1)^{1/2} \delta_k^{1/2} \tag{34}$$

Theorem E.2 (DDPM with L^2-accurate score estimate)**.** *Let $0 < \varepsilon_\chi, \varepsilon_{\mathrm{TV}}, \delta < \frac{1}{2}$. Suppose that Assumption 6 for a sufficiently small value of c that R_0 is such that $R\left(\frac{c\varepsilon_{\mathrm{TV}}^3 \delta^6 \varepsilon_\chi^{12}}{R_0^{19} d^5}\right) \le R_0$, and $R_0^2 \ge d$. Suppose one of the following cases holds.*

1. *Let $P_{\text{data}}, s(\cdot, t)$ be such that Assumption 1 holds, with $R_0^2 \ge d$. Suppose that*
$$\varepsilon_\sigma = O\left(\frac{\varepsilon_{\mathrm{TV}} \delta^{5/2} \varepsilon_\chi^{11/2}}{B^{9/4}}\right),$$
where $B = R_0^4 d \ln\left(\frac{T}{\delta}\right) \ln\left(\frac{R_0 d}{\delta \varepsilon_{\mathrm{TV}} \varepsilon_\chi}\right)$, and we run (5) starting from p_{prior} for time $T = \ln\left(\frac{16 R_0^2}{\varepsilon_\chi^2}\right)$, $N = O\left(\frac{B\left(T + \frac{1}{\delta^2}\right)}{\varepsilon_\chi^2}\right)$ steps with step sizes satisfying $h_k = O\left(\frac{\varepsilon_\chi^2}{B \max\{T - t_k, (T-t_k)^{-3}\}}\right)$.

2. *Let $P_{\text{data}}, s(\cdot, t)$ be such that Assumptions 1 and 3 hold, with $C \ge R_0^2$. Suppose*
$$\varepsilon_\sigma = O\left(\frac{\varepsilon_{\mathrm{TV}} \varepsilon_\chi^3}{T^{5/2} B}\right),$$
where $B = C^2 d \ln\left(\frac{T}{\delta}\right) \ln\left(\frac{R_0 d}{\delta \varepsilon_{\mathrm{TV}} \varepsilon_\chi}\right)$, and we run (5) starting from p_{prior} for time $T = \ln\left(\frac{16 R_0^2}{\varepsilon_\chi^2}\right)$, $N = O\left(\frac{B\left(T + \ln\left(\frac{1}{\delta}\right)\right)}{\varepsilon_\chi^2}\right)$ steps with step sizes satisfying $h_k = O\left(\frac{\varepsilon_\chi^2}{B \max\{T - t_k, (T-t_k)^{-1}\}}\right)$.

Then the resulting distribution q_{t_N} is such that q_{t_N} is $\varepsilon_{\mathrm{TV}}$-far in TV distance from a distribution $\overline{q}_{t_N}$, where $\overline{q}_{t_N}$ satisfies $\chi^2(\overline{q}_{t_N} || p_{t_N}) \le \varepsilon_\chi^2$. In particular, taking $\varepsilon_\chi = \varepsilon_{\mathrm{TV}}$, we have $\mathrm{TV}(q_T, P_{\text{data}}) \le 2\varepsilon_{\mathrm{TV}}$.

Note that the condition on R can be satisfied if $R(\varepsilon) = o(R^{-1/19})$ (no effort has been made to optimize the exponent).

Proof. We invoke Lemma C.2 for a ε_K to be chosen, to obtain a distribution $\widetilde{P}_0$ on $B_{R_0}(0)$, where $R_0 \ge R(\varepsilon_K/8)$. Let $B = R_0^4 d \ln\left(\frac{T}{\delta}\right) \ln\left(\frac{R_0}{\delta \varepsilon_K}\right)$ and $B = C^2 d \ln\left(\frac{T}{\delta}\right) \ln\left(\frac{R_0}{\delta \varepsilon_K}\right)$ in case 1 and case 2, respectively; our choice of $\varepsilon_K = O\left(\frac{\varepsilon_{\mathrm{TV}}^2 \delta^6}{n^2 R_0^6}\right)$ will give the definition of B in the theorem statement. In the following, we define $\widetilde{p}_t$ with $\widetilde{P}_0$, rather than P_{data}, as the initial distribution. Note that since $\mathrm{TV}(P_{\text{data}}, \widetilde{P}_0) \le \sqrt{\varepsilon_K} = o(\varepsilon_{\mathrm{TV}})$ (and the same holds for their evolutions under (1)), it suffices to consider convergence to $\widetilde{p}_\delta$.

We first define the bad sets where the error in the score estimate is large,

$$B_t := \{\|\nabla \ln \widetilde{p}_t(x) - s(x,t)\| > \varepsilon_{\infty,t}\} \tag{35}$$

for some $\varepsilon_{\infty,t}$ to be chosen.

Given $t \ge 0$, let $t_- = t_k$ where k is such that $t \in [t_k, t_{k+1})$. Given bad sets B_t, define the interpolated process on $[t_k, t_{k+1})$ by

$$d\overline{z}_t = g(T-t)^2 \left(\frac{1}{2} z_t + b(z_-, T - t_-)\right) dt + g(T-t)\, dw_t \tag{36}$$

$$\text{where } b(z,t) = \begin{cases} s(z,t), & z \notin B_t \\ \nabla \ln \widetilde{p}_t(z), & z \in B_t \end{cases}.$$

In other words, simulate the reverse SDE using the score estimate as long as the point is in the good set at the previous discretization timepoint t_k, and otherwise use the actual gradient $\nabla \ln p_t$. Let $\overline{q}_t$ denote the distribution of $\overline{z}_t$ when $\overline{z}_0 \sim q_0$. Note that this process is defined only for purposes of analysis, as we do not have access to $\nabla \ln p_t$. As before, we let denote q_t the distribution of z_t defined by (12).

We can couple this process with the exponential integrator (5) using s so that as long as $x_{t_m} \notin B_{T-t_m}$, the processes agree, thus satisfying condition 1 of Theorem E.1.

Then by Lemma D.3, substituting the bound of Lemma D.2 for the L^2 score error,

$$\widetilde{P}_t^{(0)}(B_t) = \varepsilon_K + \frac{4}{\varepsilon_{\infty,t}^2}\left(\varepsilon_t^2 + O\left(\frac{\varepsilon_K\left(d + \ln\left(\frac{1}{\varepsilon_K}\right)\right)}{\sigma_t^2}\right)\right),$$

Then by choice of h_k and either Corollary B.11 or B.12, when $\int_0^{t_n} \varepsilon_t^2\, dt = O(1)$,

$$\begin{aligned}\chi^2(\overline{q}_{t_k}||p_{t_k}) &= e^{\varepsilon}\chi^2(q_0||p_0) + \varepsilon + e^{\varepsilon}\int_0^{t_n} \varepsilon_{\infty,T-t}^2\, dt \\ &\leq 2\chi^2(q_0||p_0) + O(1),\end{aligned} \tag{37}$$

where $\varepsilon = \frac{\varepsilon_\chi^2}{4}$. For $\chi^2(\overline{q}_{t_k}||p_{t_k})$ to be bounded by ε_χ^2, it suffices for the terms in (37) to be bounded by $\frac{\varepsilon_\chi^2}{2}, \frac{\varepsilon_\chi^2}{4}, \frac{\varepsilon_\chi^2}{4}$; this is implied by

$$T = \ln\left(\frac{16R^2}{\varepsilon_\chi^2}\right) \text{ by Lemma B.14}$$

$$\int_0^{t_n} \varepsilon_{\infty,T-t}^2\, dt = O(\varepsilon_\chi^2). \tag{38}$$

By Theorem E.1,

$$\begin{aligned}\mathrm{TV}(q_{t_n}, \overline{q}_{t_n}) &\leq \sum_{k=0}^{n-1}(1 + \chi^2(\overline{q}_{t_k}||p_{t_k}))^{1/2} P(B_{t_k})^{1/2} \\ &\leq \sum_{k=0}^{n-1}\left(2\chi^2(q_0||p_0)^{1/2} + O(1)\right)\delta_t^{1/2} \qquad (39)\\ &= O\left(\sum_{k=0}^{n-1}\frac{\varepsilon_{t_k}}{\varepsilon_{\infty,t_k}} + \sqrt{\varepsilon_K}\left(1 + \frac{\sqrt{d + \ln(1/\varepsilon_K)}}{\varepsilon_{\infty,t_k}\sigma_{T-t_k}}\right)\right). \qquad (40)\end{aligned}$$

For this to be bounded by $\varepsilon_{\mathrm{TV}}$, it suffices for

$$\sum_{k=0}^{n-1}\frac{\varepsilon_t}{\varepsilon_{\infty,t}} = O(\varepsilon_{\mathrm{TV}}) \tag{41}$$

$$\varepsilon_K = O\left(\frac{\min_k \varepsilon_{t_k}^2 \sigma_{T-t_k}^2}{d + \ln(1/\varepsilon_K)}\right). \tag{42}$$

We bound (42) crudely, as the dependence on ε_K will be logarithmic. Using $\varepsilon_{t_k}^2 = \varepsilon_\sigma^2/\sigma_{t_k}^4$, it suffices that

$$\varepsilon_K = O\left(\frac{\varepsilon_\sigma^2}{d + \ln(1/\varepsilon_K)}\right). \tag{43}$$

We will return to this after deriving a condition on ε_σ. It remains to bound (38) and (41). We break up the timepoints depending on whether $T - t > 1$. Let

$$(t_0, t_1, \ldots, t_N) = (t_0, \ldots, t_{n^{\text{coarse}}-1}, t_0', \ldots, t_{n^{\text{fine}}}')$$

and $u_k = T - t'_k$, where $t_{n^{\text{coarse}}-1} \le T - 1 \le t'_1$. Let $h'_k = t'_{k+1} - t'_k$. Note the "fine" timepoints will be closer together than the "coarse" timepoints. We break up the integral (38) and the sum (41) into the parts involving the coarse and fine timepoints. For (38), it suffices to have

$$(38),\ \text{coarse:}\quad \int_0^{t'_0} \varepsilon^2_{\infty,T-t}\,dt \le T \max_{0\le k\le n^{\text{coarse}}} \varepsilon^2_{\infty,T-t_k} = O(\varepsilon^2_\chi)$$

so it suffices to take $\varepsilon^2_{\infty,T-t_k} \asymp \frac{\varepsilon^2_\chi}{T}$. Let $\alpha = 3$ in case 1 and $\alpha = 1$ in case 2. For the fine part, recalling our choice of h'_k, it suffices to have (note we can redefine $\varepsilon_t = \varepsilon_{t_k}$ when $t \in [t_k, t_{k+1})$ without any harm)

$$\begin{aligned}
(38),\ \text{fine:}\quad \int_{t'_0}^{t'_{n^{\text{fine}}}} \varepsilon^2_{\infty,T-t}\,dt &= \sum_{k=0}^{n^{\text{fine}}-1} h'_k \varepsilon^2_{\infty,T-t'_k} = O(\varepsilon^2_\chi)\\
\Longleftarrow \sum_{k=0}^{n^{\text{fine}}-1} \frac{\varepsilon^2_\chi u_k^\alpha}{B}\varepsilon^2_{\infty,u_k} &= O(\varepsilon^2_\chi)\\
\Longleftarrow \sum_{k=0}^{n^{\text{fine}}-1} \frac{u_k^\alpha \varepsilon^2_{\infty,u_k}}{B} &= O(1).
\end{aligned} \tag{44}$$

For (41), it suffices to have

$$\begin{aligned}
(41),\ \text{coarse:}\quad \sum_{k=0}^{n^{\text{coarse}}-1} \frac{\varepsilon_{T-t_k}}{\varepsilon_{\infty,T-t_k}} &\asymp n^{\text{coarse}} \frac{\varepsilon_\sigma}{\varepsilon_\chi/\sqrt{T}} = O(\varepsilon_{\text{TV}})\\
\Longleftarrow \varepsilon_\sigma &= O\left(\frac{\varepsilon_{\text{TV}}\varepsilon_\chi}{n^{\text{coarse}}\sqrt{T}}\right)
\end{aligned} \tag{45}$$

and

$$(41),\ \text{fine:}\quad \sum_{k=0}^{n^{\text{fine}}-1} \frac{\varepsilon_{u_k}}{\varepsilon_{\infty,u_k}} \asymp \sum_{k=0}^{n^{\text{fine}}-1} \frac{\varepsilon_\sigma}{u_k \varepsilon_{\infty,u_k}} = O(\varepsilon_{\text{TV}}). \tag{46}$$

Note that in light of the required step sizes, we can take $n^{\text{coarse}} \asymp \frac{T^2 B}{\varepsilon^2_\chi}$. Considering the equality case of Hölder's inequality on (44)$^{1/3}$(46)$^{2/3}$ suggests that we take

$$\varepsilon_\sigma \asymp \frac{\varepsilon_{\text{TV}} B^{1/2}}{\left(\sum_{k=0}^{n^{\text{fine}}-1} u_k^{\frac{\alpha-2}{3}}\right)^{3/2}} \tag{47}$$

$$\varepsilon_{\infty,u_k} \asymp \frac{B^{1/2}}{u_k^{\frac{\alpha+1}{3}} \left(\sum_{k=0}^{n^{\text{fine}}-1} u_k^{\frac{\alpha-2}{3}}\right)^{1/2}} \tag{48}$$

Note that the number of steps needed in the fine part is $O\left(\frac{B}{\varepsilon^2_\chi \delta^2}\right)$ in the first case and $O\left(\frac{B}{\varepsilon^2_\chi}\right)\ln\left(\frac{1}{\delta}\right)$ in the second case. We can check that (47) and (48) make (44) and (46) satisfied.

Finally, we calculate the denominator for ε_σ. In case 1, note that starting from $T - t'_0 = O(1)$ and taking steps of size $h'_k \asymp \frac{\varepsilon^2_\chi}{B(T-t'_k)^3}$, it takes $n^{\text{fine}} = \Theta\left(\frac{B}{\varepsilon^2_\chi \delta^2}\right)$ steps to reach $T - t = \delta$.

$$u_k = T - t'_k = \left(1 + \Theta\left(\frac{k\varepsilon^2_\chi}{B}\right)\right)^{-\frac{1}{2}}$$

$$\sum_{k=0}^{n^{\text{fine}}-1} u_k^{1/3} \asymp \sum_{k=0}^{n^{\text{fine}}-1} \left(1 + \Theta\left(\frac{k\varepsilon^2_\chi}{B}\right)\right)^{-\frac{1}{6}} \asymp \frac{B}{\varepsilon^2_\chi}(n^{\text{fine}})^{\frac{5}{6}} \asymp \left(\frac{B}{\varepsilon^2_\chi}\right)^{11/6} \frac{1}{\delta^{5/3}}.$$

Then we obtain

$$\varepsilon_\sigma \asymp \varepsilon_{\mathrm{TV}} B^{1/2} \frac{\varepsilon_\chi^{11/2}}{B^{11/4}} \delta^{5/2} = \frac{\varepsilon_{\mathrm{TV}} \delta^{5/2} \varepsilon_\chi^{11/2}}{B^{9/4}}.$$

In case 1, our requirement is

$$\varepsilon_\sigma \asymp O\left(\frac{\varepsilon_{\mathrm{TV}} \delta^{5/2} \varepsilon_\chi^{11/2}}{B^{9/4}} \wedge \frac{\varepsilon_{\mathrm{TV}} \varepsilon_\chi^3}{T^{5/2} B} \right),$$

but note that the first bound is more stringent. Now, returning to (43), we see that it suffices to take $\varepsilon_K = O\left(\frac{1}{d} \left(\frac{\varepsilon_{\mathrm{TV}} \delta^{5/2} \varepsilon_\chi^{11/2}}{R_0^9 d^{9/4}} \right)^{2+\beta} \right)$ for any $\beta > 0$ (this will "solve" the $\log(1/\varepsilon_K)$ appearing in B.)

In case 2, we have instead $u_k = \exp\left(-\Theta\left(\frac{\varepsilon_\chi^2}{B} k\right)\right)$ so

$$\varepsilon_\sigma \asymp \varepsilon_{\mathrm{TV}} B^{1/2} \left(\frac{\varepsilon_\chi^2}{B} \right)^{3/2} = \frac{\varepsilon_{\mathrm{TV}} \varepsilon_\chi^3}{B}. \qquad \square$$

Theorem E.3 (TV error for DDPM with L^2-accurate score estimate and smoothness)**.** *Let $0 < \varepsilon_{\mathrm{TV}} < \frac{1}{2}$. Suppose that Assumption 4 and 6 for a sufficiently small value of c that R_0 is such that $R\left(\frac{c\varepsilon_{\mathrm{TV}}^{15}}{R_0^{31} d^5 L^{12}}\right) \le R_0$, and $R_0^2 \ge \max\left\{d, \mathbb{E}_{P_{data}}\left[\|X\|^2\right]\right\}$, and one of the following cases holds.*

1. *Let $P_{\text{data}}, s(\cdot, t)$ be such that Assumption 1 holds. Suppose that*

$$\varepsilon_\sigma = O\left(\frac{\varepsilon_{\mathrm{TV}}^{11.5}}{B^{9/4} R_0^5 L^5} \right),$$

where $B = R_0^4 d \ln\left(\frac{T R_0^2 L^2}{\varepsilon_{\mathrm{TV}}^2}\right) \ln\left(\frac{R_0^3 d L^2}{\varepsilon_{\mathrm{TV}}^3 \varepsilon_\chi}\right)$, and we run (5) starting from p_{prior} for time $T = \ln\left(\frac{16 R_0^2}{\varepsilon_{\mathrm{TV}}^2}\right)$, $N = O\left(\frac{B\left(T + \left(\frac{R_0 L}{\varepsilon_{\mathrm{TV}}}\right)^4\right)}{\varepsilon_{\mathrm{TV}}^2} \right)$ steps with step sizes satisfying $h_k = O\left(\frac{\varepsilon_\chi^2}{B \max\{T - t_k, (T - t_k)^{-3}\}}\right)$.

2. *Let $P_{\text{data}}, s(\cdot, t)$ be such that Assumptions 1 and 3 hold, with $C \ge R_0^2$. Suppose*

$$\varepsilon_\sigma = O\left(\frac{\varepsilon_{\mathrm{TV}}^4}{T^{5/2} B} \right),$$

where $B = C^2 d \ln\left(\frac{T R_0^2 L^2}{\varepsilon_{\mathrm{TV}}^2}\right) \ln\left(\frac{R_0^3 d L^2}{\varepsilon_{\mathrm{TV}}^4}\right)$, and we run (5) starting from p_{prior} for time $T = \ln\left(\frac{16 R_0^2}{\varepsilon_{\mathrm{TV}}^2}\right)$, $N = O\left(\frac{B\left(T + \ln\left(\frac{R_0 L}{\varepsilon_{\mathrm{TV}}}\right)\right)}{\varepsilon_{\mathrm{TV}}^2} \right)$ steps with step sizes satisfying $h_k = O\left(\frac{\varepsilon_\chi^2}{B \max\{T - t_k, (T - t_k)^{-1}\}}\right)$.

Then the resulting distribution q_{t_N} is such that q_{t_N} is $\varepsilon_{\mathrm{TV}}$-far in TV distance from the data distribution P_{data}.

Proof. With the result of Theorem E.2, we see that $\mathrm{TV}(q_{t_N}, p_{t_N}) \le 2\varepsilon_{\mathrm{TV}}$. Now by Lemma D.4, if we further assume

$$\delta = O\left(\frac{\varepsilon_{\mathrm{TV}}^2}{R_0^2 L^2} \right),$$

then $\mathrm{TV}(p_{t_N}, P_{\text{data}}) \le \varepsilon_{\mathrm{TV}}$. We conclude the proof by triangle inequality and replacing the δ-dependence with $O(\frac{\varepsilon_{\mathrm{TV}}^2}{R_0^2 L^2})$ in the previous theorem. $\square$

Proof of Theorem 3.3. If P_{data} is subexponential with a fixed constant, note that Assumption 6 holds with $R(\varepsilon) = O\left(\ln\left(\frac{1}{\varepsilon}\right)\right)$ and hence R_0 is logarithmic in all parameters. $\square$

E.2 Wasserstein error guarantees

Proof of Theorem 3.1. Suppose $T - t_N = \delta$. Note by (3) that $M_{m_\delta^{-1}\#}\widetilde{p}_\delta$ has the same distribution as $\widetilde{x}_0 + m_\delta^{-1}\sigma_\delta z$, where $\widetilde{x}_0 \sim \widetilde{P}_0$ and $z \sim N(0, I_d)$. Then $W_2(\widetilde{p}_0, M_{m_\delta^{-1}\sharp}\widetilde{p}_\delta) \leq m_\delta^{-1}\sigma_\delta\sqrt{d} \leq \sqrt{e^\delta - 1} \leq \sqrt{2\delta d}$ (for $\delta \leq 1$). Choosing $\delta = \frac{\varepsilon_{\mathrm{W}}^2}{2d}$, we see by Theorem E.2 it suffices to take

$$\varepsilon_\sigma = O\left(\frac{\varepsilon_{\mathrm{TV}}^{13/2}(\varepsilon_{\mathrm{W}}^2/d)^{5/2}}{\left(R^4 d \ln\left(\frac{T}{\delta}\right)\ln\left(\frac{RN}{\delta\varepsilon_{\mathrm{W}}}\right)\right)^{9/4}}\right).$$

Simplifying gives $\varepsilon_\sigma = \widetilde{o}\left(\frac{\varepsilon_{\mathrm{TV}}^{6.5}\varepsilon_{\mathrm{W}}^5}{R^9 d^{4.75}}\right)$. If Assumption 3 also holds, then it suffices to take

$$\varepsilon_\sigma = O\left(\frac{\varepsilon_{\mathrm{TV}}^4}{T^{5/2}C^2 d \ln\left(\frac{T}{\delta}\right)\ln\left(\frac{RN}{\delta\varepsilon_{\mathrm{W}}}\right)}\right).$$

Simplifying gives $\varepsilon_\sigma = \widetilde{o}\left(\frac{\varepsilon_{\mathrm{TV}}^4}{C^2 d}\right)$. □

Proof of Theorem 3.2. To obtain purely Wasserstein error guarantees, we include an extra step of replacing any sample $z_{t_N} \sim q_{t_N}$ falling outside $B_R(0)$ by 0. Suppose $T - t_N = \delta$. Let $\widehat{q}_{t_N}$ be the resulting distribution. Then as above,

$$\begin{aligned} W_2(\widetilde{p}_0, \widehat{q}_{t_N}) &\leq W_2(\widetilde{p}_0, \widetilde{p}_\delta) + W_2(\widetilde{p}_\delta, \widehat{q}_{t_N}) \\ &\leq \sqrt{2\delta d} + W_2(\widetilde{p}_\delta, \widehat{q}_{t_N}). \end{aligned}$$

We choose $\delta = \frac{\varepsilon_{\mathrm{W}}^2}{8d}$ so the first term is $\leq \frac{\varepsilon_{\mathrm{W}}}{2}$. It suffices to bound the second term $W_2(\widetilde{p}_\delta, \widehat{q}_{t_N})$ also by $\frac{\varepsilon_{\mathrm{W}}}{2}$. We bound it in terms of $\mathrm{TV}(\widetilde{p}_\delta, \widehat{q}_{t_N})$ using the fact that $\widehat{q}_{t_N}$ is supported on $B_R(0)$ and using a Gaussian tail calculation for $\widetilde{p}_\delta$. Consider a coupling of $x_{t_N} = \widetilde{x}_\delta \sim \widetilde{p}_\delta$ and $\widehat{z}_{t_N} \sim \widehat{q}_{t_N}$ such that $x_\delta \neq \widehat{z}_{t_N}$ with probability $\varepsilon_{\mathrm{TV}}$. Express $\widetilde{x}_\delta = m_\delta\widetilde{x}_0 + \sigma_\delta\xi$ where $\widetilde{x}_0 \sim \widetilde{p}_0$. Now

$$\begin{aligned} \mathbb{E}[\|\widetilde{x}_\delta - \widehat{z}_{t_N}\|^2] &\leq \sup_{P(A)\leq\varepsilon_{\mathrm{TV}}} 2\left(\mathbb{E}[\|m_\delta\widetilde{x}_0 - z_{t_N}\|^2 \mathbb{1}_A] + \sigma_\delta^2\mathbb{E}[\|\xi\|^2 \mathbb{1}_A]\right) \\ &= 2\left(4R^2\varepsilon_{\mathrm{TV}} + \sigma_\delta^2\varepsilon_{\mathrm{TV}} \cdot O\left(d + \ln\left(\frac{1}{\varepsilon_{\mathrm{TV}}}\right)\right)\right), \end{aligned}$$

where the bound on the second term uses Lemma D.6. Using $R^2 \geq d$, we see that it suffices to choose $\varepsilon_{\mathrm{TV}} = O\left(\frac{\varepsilon_{\mathrm{W}}^2}{R^2}\right)$ for appropriate choice of constants. By Theorem E.2, it suffices to take

$$\varepsilon_\sigma = O\left(\frac{(\varepsilon_{\mathrm{W}}^2/R^2)^{13/2}\left(\varepsilon_{\mathrm{W}}^2/d\right)^{5/2}}{\left(R^4 d \ln\left(\frac{T}{\delta}\right)\ln\left(\frac{RN}{\delta\varepsilon_{\mathrm{W}}}\right)\right)^{9/4}}\right).$$

Simplifying gives $\widetilde{o}\left(\frac{\varepsilon_{\mathrm{W}}^{18}}{R^{22}d^{4.75}}\right)$.

In case 2, it suffices to take

$$\varepsilon_\sigma = O\left(\frac{(\varepsilon_{\mathrm{W}}^2/R^2)^4}{T^{5/2}(C^2 d \ln\left(\frac{T}{\delta}\right)\ln\left(\frac{RN}{\delta\varepsilon_{\mathrm{W}}}\right))}\right).$$

Simplifying gives $\varepsilon_\sigma = \widetilde{o}\left(\frac{\varepsilon_{\mathrm{W}}^8}{C^2R^8 d}\right)$. □

F High-probability bound on the Hessian

In this section we obtain a high-probability bound on the Hessian of $\ln \widetilde{p}_t$, i.e., the Jacobian of the score function.

To see why we expect Hessian to usually be smaller than the worst-case bound given by Lemma B.13, note that we can express (27) and (28) as

$$\nabla \ln(\mu * \varphi_{\sigma^2}(y)) = -\frac{1}{\sigma^2}\mathbb{E}[Y - X|Y = y] \tag{49}$$

$$\nabla^2 \ln(\mu * \varphi_{\sigma^2}(y)) = \frac{1}{\sigma^4}\operatorname{Cov}[Y - X|Y = y] - \frac{1}{\sigma^2} I_d \tag{50}$$

where $X \sim \mu$ and $Y = X + \sigma\xi$, $\xi \sim N(0, I_d)$. We expect that the random variable $Y - X$ is distributed as $N(0, \sigma^2 I_d)$, which suggests that the covariance (50) may be bounded by $\frac{1}{\sigma^2}$ rather than $\frac{1}{\sigma}$ with high probability. Indeed, we can easily construct an example where the worst case of Lemma B.13 is attained—for example, $\mu = \frac{1}{2}(\delta_{-v} + \delta_v)$ for $\|v\|_2 = R$, at $x = 0$—but this point has exponentially small probability density under $\mu * \varphi_{\sigma^2}$.

The following lemma uses a ε-net argument to bound the operator norm of the variance of a conditional distribution, with high probability.

Lemma F.1. *Suppose X is a $\mathbb{R}^d$-valued random variable over the probability space $(\Omega, \mathcal{G}, P)$, and $\mathcal{F} \subseteq \mathcal{G}$ is a σ-subalgebra. If X is subgaussian, then*

$$\mathbb{P}\left(\mathbb{E}\left[\left\|XX^\top\right\| |\mathcal{F}\right] \geq 2\|X\|_{\psi_2}^2 \ln\left(\frac{2 \cdot 5^d}{\varepsilon}\right)\right) \leq \varepsilon.$$

Proof. By Jensen's inequality and Markov's inequality, for any $v \in \mathbb{S}^{d-1}$,

$$\begin{aligned}
\mathbb{P}\left(\mathbb{E}[v^\top XX^\top v|\mathcal{F}] \geq \lambda^2\right) &= \mathbb{P}\left(e^{\mathbb{E}[v^\top XX^\top v|\mathcal{F}]/c^2} \geq e^{\lambda^2/c^2}\right) \\
&\leq \mathbb{P}\left(\mathbb{E}\left[e^{\langle X, v\rangle^2/c^2}|\mathcal{F}\right] \geq e^{\lambda^2/c^2}\right) \\
&\leq \frac{\mathbb{E}\left[\mathbb{E}[e^{\langle X, v\rangle^2/c^2}|\mathcal{F}]\right]}{e^{\lambda^2/c^2}} = \frac{\mathbb{E}\left[e^{\langle X, v\rangle^2/c^2}\right]}{e^{\lambda^2/c^2}} \leq 2e^{-\lambda^2/\|X\|_{\psi_2}},
\end{aligned}$$

where the last inequality follows from taking $c = \|X\|_{\psi_2}$. Now take a $\frac{1}{2}$-net $\mathcal{N}$ of $\mathbb{S}^{d-1}$ of size $\leq 5^d$ [Ver18, Cor. 4.2.13]. By a union bound,

$$\mathbb{P}\left(\exists v \in \mathcal{N} : \mathbb{E}[v^\top XX^\top v|\mathcal{F}] \geq \lambda^2\right) \leq 5^d \cdot 2 \cdot e^{-\lambda^2/\|X\|_{\psi_2}^2} = \varepsilon$$

when we take $\lambda = \|X\|_{\psi_2}\sqrt{\ln\left(\frac{2 \cdot 5^d}{\varepsilon}\right)}$. By [Ver18, Lemma 4.4.1], the operator norm can be bounded by the norm on an ε-net,

$$\|A\| \leq 2\sup_{v \in A}\|\langle A, v\rangle\| = 2\sup_{v \in A}|v^\top Av|.$$

where the second inequality holds when A is symmetric. The result follows from applying this to $\mathbb{E}[v^\top XX^\top v|\mathcal{F}]$. □

From this we obtain the desired high-probability bound.

Lemma F.2. *There is a universal constant C such that the following holds. For any starting distribution $\widetilde{P}_0$, letting $\widetilde{P}_t$ be the law of the DDPM process* (1) *at time t, we have*

$$\widetilde{P}_t\left(\left\|\nabla^2 \ln \widetilde{p}_t(x)\right\| \leq \frac{C(d + \ln\left(\frac{1}{\varepsilon}\right))}{\sigma_t^2}\right) \geq 1 - \varepsilon.$$

Note that there is no dependence on the radius.

Proof. Apply (50) with $\mu = M_{m_t\sharp}\widetilde{P}_0$ to obtain $\nabla^2 \ln \widetilde{p}_t$. Noting that $Y - X \sim N(0, \sigma^2 I_d)$ is subgaussian with $\|Y - X\|_{\psi_2} \leq C_2\sigma$ for some universal constant C_2, the result follows from Lemma F.1. □