Analysis of Score-based Generative Models

by

Yixin Tan

Department of Mathematics
Duke University

Defense Date: March 20, 2024

Approved:

Jianfeng Lu, Supervisor

Xiuyuan Cheng

Jonathan C. Mattingly

James H. Nolen

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Mathematics
in the Graduate School of Duke University
2024

ABSTRACT

Analysis of Score-based Generative Models

by

Yixin Tan

Department of Mathematics
Duke University

Defense Date: March 20, 2024

Approved:

Jianfeng Lu, Supervisor

Xiuyuan Cheng

Jonathan C. Mattingly

James H. Nolen

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Mathematics in the Graduate School of Duke University
2024

Abstract

In this thesis, we study the convergence of diffusion models and related flow-based methods, which are highly successful approaches for learning a probability distribution from data and generating further samples. For diffusion models, we established the first convergence result applying to data distributions satisfying the log-sobolev inequality without suffering the curse of dimensionality. Our analysis gives theoretical grounding to the observation that an annealed procedure is required in practice to generate good samples, as our proof depends essentially on using annealing to obtain a warm start at each step. Moreover, we show that a predictor-corrector algorithm gives better convergence than using either portion alone. Then we generalized the results to any distribution with bounded 2nd moment, relying only on a L^2-accurate score estimates, with polynomial dependence on all parameters and no reliance on smoothness or functional inequalities. We also provide a theoretical guarantee of generating data distribution by a progressive flow model, the so-called JKO flow model, which implements the Jordan-Kinderleherer-Otto (JKO) scheme in a normalizing flow network. Leveraging the exponential convergence of the proximal gradient descent (GD) in Wasserstein space, we prove the Kullback-Leibler (KL) guarantee of data generation by a JKO flow model where the assumption on data density is merely a finite second moment.

Contents

List of Figures

Acknowledgements

I would like to express my deepest gratitude to my advisor, Professor Jianfeng Lu. His guidance and support throughout my graduate studies have been invaluable, and I am grateful for his expertise and encouragement.

I am also grateful to my research collaborators, including Xiuyuan Cheng, Wei Deng, Haque Ishfaq, Yao Xie, Pan Xu and Hao Zhang, for their valuable contributions and collaboration. I appreciate the opportunity to have learned alongside them.

Many thanks to the committee members of my preliminary exam and thesis defense: Prof. Xiuyuan Cheng, Prof. Jonathan Mattingly, Prof. James Nolen and Prof. David Dunson. Their feedback and insights have been instrumental in advancing my research.

I would like to extend my sincere thanks to the staff in the department, including Julia Gruhot, Yunliang Yu, Laurie Triggiano, etc. Their dedication to creating a supportive and comfortable research environment is deeply appreciated.

Finally, I would like to acknowledge the unwavering support of my family, girlfriend, and friends. Their encouragement and love have been my rock throughout my PhD journey.

1. Introduction

Generative models, from generative adversarial networks (GAN) (I. J. Goodfellow et al., 2014; Gulrajani et al., 2017; Isola et al., 2017) and variational auto-encoder (VAE) (Kingma & Welling, 2013, 2019a) to normalizing flow (Kobyzev et al., 2020), have achieved many successes in applications and have become a central topic in deep learning. More recently, diffusion models (Ho et al., 2020; Song & Ermon, 2019; Song, Sohl-Dickstein, et al., 2021) and closely related flow-based models (Albergo et al., 2023; Albergo & Vanden-Eijnden, 2023; Fan et al., 2022; Lipman et al., 2023; Xu et al., 2023b) have drawn much research attention, given their state-of-the-art performance in image generations. In this thesis, we study the convergence of different generative models.

1.1 Convergence of Diffusion Models

Score-based generative modeling (SGM) is a highly successful approach for learning a probability distribution from data and generating further samples.We prove the first polynomial convergence guarantees for the core mechanic behind SGM: drawing samples from a probability density p given a score estimate (an estimate of $\nabla \ln p$) that is accurate in $L^2(p)$. Compared to previous works, we do not incur error that grows exponentially in time or that suffers from a curse of dimensionality. Our guarantee works for any smooth distribution and depends polynomially on its log-Sobolev constant. Using our guarantee, we give a theoretical analysis of score-based generative modeling, which transforms white-noise input into samples from a learned data distribution given score estimates at different noise scales. Our analysis gives theoretical grounding to the observation that an annealed procedure is required in practice to generate good samples, as our proof depends essentially on using annealing to obtain a warm start at each step. Moreover, we show that a predictor-corrector algorithm gives better convergence than using either portion alone. We also extend our results to 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.

Chapter 2 and 3 are based on (Lee et al., 2022) and (Lee et al., 2023), which are joint works with Prof. Holden Lee and Prof. Jianfeng Lu.

1.2 Convergence of flow-based generative models via JKO scheme

Flow-based generative models enjoy certain advantages in computing the data generation and the likelihood, and have recently shown competitive empirical performance. Compared to the accumulating theoretical studies on related score-based diffusion models, analysis of flow-based models, which are deterministic in both forward (data-to-noise) and reverse (noise-to-data) directions, remain sparse. In this paper, we provide a theoretical guarantee of generating data distribution by a progressive flow model, the so-called JKO flow model, which implements the Jordan-Kinderleherer-Otto (JKO) scheme in a normalizing flow network. Leveraging the exponential convergence of the proximal gradient descent (GD) in Wasserstein space, we prove the Kullback-Leibler (KL) guarantee of data generation by a JKO flow model to be $O(\varepsilon^2)$ when using $N \lesssim \log(1/\varepsilon)$ many JKO steps (N Residual Blocks in the flow) where ε is the error in the per-step first-order condition. The assumption on data density is merely a finite second moment, and the theory extends to data distributions without density and when there are inversion errors in the reverse process where we obtain KL-$\mathcal{W}_2$ mixed error guarantees. The non-asymptotic convergence rate of the JKO-type $\mathcal{W}_2$-proximal GD is proved for a general class of convex objective functionals that includes the KL divergence as a special case, which can be of independent interest.

Chapter 4 is based on (Cheng et al., 2023), which is a joint work with Prof. Xiuyuan Cheng, Prof. Jianfeng Lu and Prof. Yao Xie.

2. Convergence for SGMs with LSI

2.1 Background and Overview

A key task in machine learning is to learn a probability distribution from data, in a way that allows efficient generation of additional samples from the learned distribution. Score-based generative modeling (SGM) is one empirically successful approach that *implicitly* learns the probability distribution by learning how to transform white noise into the data distribution, and gives state-of-the-art performance for generating images and audio (Dathathri et al., 2019; Grathwohl et al., 2019; Jing et al., 2022; Meng et al., 2021; Song, Durkan, et al., 2021; Song & Ermon, 2019, 2020; Song, Shen, et al., 2021; Song, Sohl-Dickstein, et al., 2020). It also yields a conditional generation process for inverse problems (Dhariwal & Nichol, 2021). The basic idea behind score-based generative modeling is to first estimate the score function from data (Song, Garg, et al., 2020) and then to sample the distribution based on the learned score function. Other approaches for generative modeling include generative adversarial networks (GANs) (Arjovsky et al., 2017; I. Goodfellow et al., 2014), normalizing flows (Dinh et al., 2016), variational autoencoders (Kingma & Welling, 2019b), and energy-based models (Zhao et al., 2016).

General framework. The *score function* of a distribution P with density p is defined as the gradient of the log-pdf, $\nabla \ln p$. Its significance arises from the fact that knowing the score function allows running a variety of sampling algorithms, based on discretizations of stochastic differential equations (SDE's), to sample from p. SGM consists of two steps: first, learning an estimate of the score function for a sequence of "noisy" versions of the data distribution P_{data}, and second, using the score function in lieu of the gradient of the log-pdf in the chosen sampling algorithm. We now describe each of these steps more precisely.

First, a method of adding noise to the data distribution is fixed; this takes the form of evolving a (forward) stochastic differential equation (SDE) starting from the data distribution. We fix a sequence of noise levels $\sigma_1 < \cdots < \sigma_N$. For $\sigma \in \{\sigma_1, \ldots, \sigma_N\}$, let the resulting distributions be P_{σ^2} and the distributions conditional on the starting data point be $P_{\sigma^2}(\cdot|x)$. Typically, σ_1 is chosen so that $P_{\sigma_1^2} \approx P_{\text{data}}$ and $P_{\sigma_N^2}$ is close to some "prior" distribution that

is easy to sample from, such as $N(0, \sigma_N^2 I_d)$. While the score $\nabla \ln p_{\sigma^2}$ cannot be estimated directly, it turns out that a de-noising objective that is equivalent to the score-matching objective can be calculated (Song & Ermon, 2019). This de-noising objective can be estimated from samples $(X, \widetilde{X})$ where $\widetilde{X} \sim P_{\sigma^2}(\cdot|x)$. The objective is represented and optimized within an expressive function class, typically neural networks, to obtain a L^2-estimate of the score, that is, $s_\theta(x, \sigma^2)$ such that

$$\mathbb{E}_{x \sim P_{\sigma^2}}[\|s_\theta(x, \sigma^2) - \nabla \ln p_{\sigma^2}(x)\|^2] \tag{2.1}$$

is small.

The reason we estimate the score function $\nabla \ln p_{\sigma^2}$ is that there are a variety of sampling algorithms—based on simulating SDE's—that can sample from p given access to $\nabla \ln p$, including Langevin Monte Carlo and Hamiltonian Monte Carlo. The second step is then to use the estimated score function $s_\theta(x, t)$ in lieu of the exact gradient in the sampling algorithm to successively obtain samples from $p_{\sigma_N^2}, \ldots, p_{\sigma_1^2}$. This sequence interpolates smoothly between the prior distribution (e.g., $N(0, \sigma_N^2 I_d)$) and the data distribution P_{data}; such an "annealing" or "homotopy" method is required in practice to generate good samples (Song, Sohl-Dickstein, et al., 2020).

Examples of SGM's. There have been several instantiations of this general approach. (Song & Ermon, 2019) add gaussian noise to the data and then use Langevin diffusion at a discrete set of noise levels $\sigma_N > \cdots > \sigma_1$ as the sampling algorithm. (Song, Sohl-Dickstein, et al., 2020) take the continuous perspective and consider a more general framework, where the forward process can be any reasonable SDE. Then a natural *reverse SDE* evolves the final distribution $p_{\sigma_N^2}$ back to the data distribution; this process can be simulated with the estimated score. They consider methods based on two different SDE's: score-matching Langevin diffusion (SMLD) based on adding Gaussian noise and denosing diffusion probabilistic models (DDPM) (Ho et al., 2020; Sohl-Dickstein et al., 2015), based on the Ornstein-Uhlenbeck process. Note that a difference with MCMC-based methods is that these SDE's are evolved for a fixed amount of time, rather than until convergence. However, they can be combined

with MCMC-based methods such as Langevin diffusion in the *predictor-corrector* approach for improved convergence. (Dockhorn et al., 2021) include Hamiltonian dynamics: they augment the state space with a velocity variable and consider a critically-damped version of the Ornstein-Uhlenbeck process. Finally, we note the work of (De Bortoli et al., 2021), who introduce the Diffusion Schrödinger Bridge method to learn a diffusion that more quickly transforms the prior into the data distribution.

We will give a general analysis framework for SGM's that applies to the algorithms in both (Song & Ermon, 2019) and (Song, Sohl-Dickstein, et al., 2020).

2.1.1 Prior work and challenges for theory

Although the literature on convergence for Langevin Monte Carlo (Cheng & Bartlett, 2018; Cheng et al., 2018; Dalalyan, 2017; Dalalyan & Karagulyan, 2019; Durmus & Moulines, 2017; Erdogdu et al., 2021; Majka et al., 2020) and related sampling algorithms is extensive, prior works mainly consider the case of exact or stochastic gradients. In contrast, by the structure of the loss function (2.1), the score function learned in SGM is only accurate in $L^2(p)$. This poses a significant challenge for analysis, as the stationary distribution of Langevin diffusion with $L^2(p)$-accurate gradient can be arbitrarily far from p (see Appendix 2.8). Hence, any analysis must be utilizing the short/medium-term convergence, while overcoming the potential issue of long-term behavior of convergence to an incorrect distribution.

(Block et al., 2020) give the first theoretical analysis of SGM, and in particular, Langevin Monte Carlo with $L^2(p)$-accurate gradients. First, they show using uniform generalization bounds that optimizing the de-noising autoencoder (DAE) objective does in fact give a $L^2(p)$-accurate score function, with sample complexity depending on the complexity of the function class. They analyze convergence of LMC in Wasserstein distance. However, the error they obtain (Theorem 13) only decreases as $\varepsilon^{1/d}$ where ε is the accuracy of the score estimate—so it suffers from the curse of dimensionality—and increases exponentially in the time that the process is run, the dimension, and the smoothness of the distribution, as in

ODE/SDE discretization arguments that do not depend on contractivity.

(De Bortoli et al., 2021) give an analysis for (Song, Sohl-Dickstein, et al., 2020) in TV distance that requires a L^∞-accurate score function and depends exponentially on the amount of time the reverse SDE is run. Although exponential dependence is bad in general, it is mollified using their Diffusion Schrödinger Bridge (DSB) approach, as it allows running for a shorter, fixed amount of time, before the forward SDE converges to the prior distribution. However, this supposes that a good solution can be found for the DSB problem, and theoretical guarantees may be difficult to obtain.

We overcome the challenges of analysis with a $L^2(p)$-accurate gradient, and give the first analysis with only polynomial dependence on running time, dimension, and smoothness of the distribution, with rates that are a fixed power of ε. Our convergence result is in TV distance. We assume only smoothness conditions and a bounded log-Sobolev constant of the data distribution, a weaker condition than the dissipativity condition required by (Block et al., 2020). We introduce a general framework for analysis of sampling algorithms given L^2-accurate gradients (score function) based on constructing a "bad set" with small measure and showing convergence of the discretized process conditioned on not hitting the bad set. We use our framework to give an end-to-end analysis for both the algorithms in (Song & Ermon, 2019) and (Song, Sohl-Dickstein, et al., 2020), and illuminate the relative performance of different methods in practice.

2.1.2 Notation and organization

Through out the chapter, $p(x) \propto e^{-V(x)}$ denotes the target distribution in $\mathbb{R}^d$ and $V : \mathbb{R}^d \to \mathbb{R}$ is referred to as the potential. We abuse notation by identifying a measure with its density when context allows. We write $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$. We use $a = O(b)$ or $b = \Omega(a)$ to indicate that $a \leqslant Cb$ for a universal constant $C > 0$. Also, we write $a = \Theta(b)$ if there are universal constants $c' > c > 0$ such that $cb \leqslant a \leqslant cb$, and the notation $\tilde{O}(\cdot)$ means it hides polylog factors in the parameters. Definite integrals without limits are taken over $\mathbb{R}^d$.

In Section 2.2 we explain our main results for Langevin Monte Carlo with $L^2(p)$-accurate score estimate and use it to derive convergence bounds for the annealed LMC method of (Song & Ermon, 2019). In Section 2.3, we give our main results for the predictor-corrector algorithms of (Song, Sohl-Dickstein, et al., 2020) based on simulating reverse SDE's. Our proofs are based on a common framework which we introduce in Section 2.4. Full proofs are in the appendix.

2.2 Results for Langevin dynamics with estimated score

Let $p(x) \propto e^{-V(x)}$ be a probability density on $\mathbb{R}^d$ such that V is C^1. Langevin diffusion with stationary distribution p is the stochastic process defined by the SDE

$$dx_t = -\nabla V(x_t)\, dt + \sqrt{2}\, dw_t,$$

where w_t is a standard Brownian Motion in $\mathbb{R}^d$. The rate of convergence to p in χ^2 and KL divergences are given by the Poincaré and log-Sobolev constants of p, respectively; see Section 2.9.1. To obtain the Langevin Monte Carlo (LMC) algorithm, we take the Euler-Murayama discretization of the SDE. We define LMC with score estimate $s(x) \approx -\nabla V(x)$ and step size h by

$$x_{(k+1)h} = x_{kh} + h \cdot s(x_{kh}) + \sqrt{2h} \cdot \xi_{kh}, \text{ where } \xi_{kh} \sim N(0, I_d). \tag{LMC-SE}$$

We make the following assumptions on the density p and the score estimate s, which we will use throughout this chapter.

Assumption 1. *p is a probability density on $\mathbb{R}^d$ such that the following hold.*

1. *$\ln p$ is C^1 and L-smooth, that is, $\nabla \ln p$ is L-Lipschitz. We assume $L \geqslant 1$.*
2. *p satisfies a log-Sobolev inequality with constant C_{LS}. We assume $C_{\mathrm{LS}} \geqslant 1$.*
3. *(Moments) $\left\|\mathbb{E}_p x\right\| \leqslant M_1$ and $\mathbb{E}_p \left\|x\right\|^2 \leqslant M_2$.*

We note that the uniform Lipschitzness assumption (1) helps ensure a unique strong solution to the Langevin diffusion, as in (Block et al., 2020). One special case where one can prove Lipschitzness for all t is when p_0 is strongly log-concave (Lee et al., 2021, Lemma 28).

Although satisfying a log-Sobolev inequality (3) is a significant assumption, it is standard for analysis of Langevin Monte Carlo (Vempala & Wibisono, 2019). It is much weaker than assumptions in previous works (Block et al., 2020), including log-concave distributions and distributions satisfying strong dissipativity, and is stable under bounded perturbations. See Section 2.9.1 for background on functional inequalities.

Assumption 2. *Let p be a given probability density on $\mathbb{R}^d$ such that $\ln p$ is C^1. The score estimate $s : \mathbb{R}^d \to \mathbb{R}^d$ satisfies the following.*

1. *s is a C^1 function that is L_s-Lipschitz. We assume $L_s \geqslant 1$.*
2. *The error in the score estimate is bounded in L^2:*

$$\|\nabla \ln p - s\|^2_{L^2(p)} = \mathbb{E}_p[\|\nabla \ln p(x) - s(x)\|^2] \leqslant \varepsilon^2.$$

2.2.1 Langevin with L^2-accurate score estimate

Our first main result gives an error bound between the sampled distribution and p, assuming L^2-accurate score function estimate.

Theorem 2.2.1 (LMC with L^2-accurate score estimate)**.** *Let $p : \mathbb{R}^d \to \mathbb{R}$ be a probability density satisfying Assumption 1(1, 2) with $L \geqslant 1$ and $s : \mathbb{R}^d \to \mathbb{R}^d$ be a score estimate satisfying Assumption 3(2). Consider the accuracy requirement in* TV *and χ^2: $0 < \varepsilon_{\mathrm{TV}} < 1$, $0 < \varepsilon_\chi < 1$, and suppose furthermore the starting distribution satisfies $\chi^2(p_0||p) \leqslant K_\chi^2$. Then if*

$$\varepsilon = O\left(\frac{\varepsilon_{\mathrm{TV}}\varepsilon_\chi^3}{dL^2 C_{\mathrm{LS}}^{5/2}(\ln(2K_\chi/\varepsilon_\chi^2) \vee K_\chi)}\right), \tag{2.2}$$

then running (LMC-SE) *with score estimate s, step size $h = \Theta\Big(\frac{\varepsilon_\chi^2}{dL^2 C_{\mathrm{LS}}}\Big)$, and time*

$$T = \Theta\Big(C_{\mathrm{LS}} \ln\big(\frac{2K_\chi}{\varepsilon_\chi^2}\big)\Big)$$

results in a distribution p_T such that p_T is $\varepsilon_{\mathrm{TV}}$-far in TV distance from a distribution $\overline{p}_T$, where $\overline{p}_T$ satisfies $\chi^2(\overline{p}_T||p) \leqslant \varepsilon_\chi^2$. In particular, taking $\varepsilon_\chi = \varepsilon_{\mathrm{TV}}$, we have the error guarantee that $\mathrm{TV}(p_T, p) \leqslant 2\varepsilon_{\mathrm{TV}}$.

Note that the error bound is only achieved when running LMC for a moderate time; this is consistent with the fact that the stationary distribution of LMC with a L^2-score estimate can be arbitrarily far from p. Note also that we need a warm start in χ^2-divergence: to obtain fixed errors $\varepsilon_{\mathrm{TV}}, \varepsilon_\chi$, the required accuracy for the score estimate is inversely proportional to K_χ. Intuitively, we must suffer from such a dependence because if the starting distribution is very far away, then there is no guarantee that $\|\nabla \ln p(x_t) - s(x_t)\|^2$ is small on average during the sampling algorithm. Finally, although we can state a result purely in terms of TV distance, we need this more precise formulation to prove a result for annealed Langevin dynamics.

2.2.2 Annealed Langevin dynamics with estimated score

In light of the warm start requirement in Theorem 2.2.1, we typically cannot directly sample from p_{data} or its approximation. Hence, (Song & Ermon, 2019) proposed using annealed Langevin dynamics: consider a sequence of noise levels $\sigma_N > \cdots > \sigma_1 \approx 0$ giving rise to a sequence of distributions $p_{\sigma_N^2}, \ldots, p_{\sigma_1^2} \approx p_{\mathrm{data}}$, where $p_{\sigma^2} = p * \varphi_{\sigma^2}$, φ_{σ^2} being the density of $N(0, \sigma^2 I_d)$. For large enough σ_N, $\varphi_{\sigma_N^2} \approx p_{\sigma_N^2}$ provides a warm start to $p_{\sigma_N^2}$. We then successively run LMC using score estimates for $p_{\sigma_k^2}$, with the approximate sample for $p_{\sigma_k^2}$ giving a warm start for $p_{\sigma_{k-1}^2}$. We obtain the following algorithm and error estimate.

Algorithm 1 Annealed Langevin dynamics with estimated score (Song & Ermon, 2019)

INPUT: Noise levels $0 \leqslant \sigma_1 < \ldots < \sigma_M$; score function estimates $s(\cdot, \sigma_m)$ (estimates of $\nabla \ln(p * \varphi_{\sigma_m^2})$), step sizes h_m, and number of steps N_m for $1 \leqslant m \leqslant M$.
Draw $x^{(M+1)} \sim N(0, \sigma_M^2 I_d)$.
for m from M to 1 **do**
  Starting from $x_0^{(m)} = x^{(m+1)}$, run (LMC-SE) with $s(x, \sigma_m)$ and step size h_m for N_m steps, and let the final sample be $x^{(m)}$.
end for
OUTPUT: Return $x^{(1)}$, approximate sample from $p * \varphi_{\sigma_1^2}$.

Theorem 2.2.2 (Annealed LMC with L^2-accurate score estimate)**.** *Let $p : \mathbb{R}^d \to \mathbb{R}$ be a probability density satisfying Assumption 1 for $M_1 = O(d)$, and let $p_{\sigma^2} := p * \varphi_{\sigma^2}$. Suppose furthermore that $\nabla \ln p_{\sigma^2}$ is L-Lipschitz for every $\sigma \geqslant 0$. Given $\sigma_{\min} > 0$, there exists a sequence*

$\sigma_{\min} = \sigma_1 < \cdots < \sigma_M$ *with* $M = O\left(\sqrt{d}\log\left(\frac{dC_{\mathrm{LS}}}{\sigma_{\min}^2}\right)\right)$ *such that for each* m, *if*

$$\left\|\nabla \ln(p_{\sigma_m^2}) - s(\cdot, \sigma_m^2)\right\|^2_{L^2(p_{\sigma_m^2})} = \mathbb{E}_{p_{\sigma_m^2}}[\left\|\nabla \ln p_{\sigma_m^2}(x) - s(x, \sigma_m^2)\right\|^2] \leqslant \varepsilon^2.$$

$$\text{with } \varepsilon := \widetilde{O}\left(\frac{\varepsilon_{\mathrm{TV}}^{4.5}}{d^{3.25}L^2C_{\mathrm{LS}}^{2.5}}\right) \tag{2.3}$$

then $x^{(1)}$ *is a sample from a distribution* q *such that* $\mathrm{TV}(q, p_{\sigma_1^2}) \leqslant \varepsilon_{\mathrm{TV}}$.

Note that we assume a score estimate with error ε at all noise scales; this corresponds to using an objective function that is a maximum of the score-matching objective over all noise levels, rather than an average over all noise levels as more commonly used in practice. However, these two losses are at most a factor of M apart.

The proof shows that the noise levels σ_k can be chosen as a geometric sequence, which matches the choice used in practice (Song & Ermon, 2020). The additional dependence on d and $\varepsilon_{\mathrm{TV}}$ in Theorem 2.2.2 compared to Theorem 2.2.1 comes from requiring a sequence of $\widetilde{O}(\sqrt{d})$ noise levels and an additional factor in χ^2-divergence we suffer at the beginning of each level m. In the next section, we will find that using a reverse SDE to evolve the samples between the noise levels—called a *predictor* step—will improve the rate and time complexity.

2.3 Results for reverse SDE's with estimated score

To improve the empirical performance of score-based generative modeling, Song, Sohl-Dickstein, et al., 2020 consider a general framework where noise is injected into a data distribution p_{data} via a forward SDE,

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

where $\tilde{x}_0 \sim \tilde{p}_0 := p_{\mathrm{data}}$. Let $\tilde{p}_t$ denote the distribution of $\tilde{x}_t$ ($\tilde{p}_t$ is used instead of p_t to distinguish with the Gaussian-convolved distribution used in Annealed Langevin dynamics as in §2.2.2). Remarkably, $\tilde{x}_t$ also satisfies a reverse-time SDE,

$$d\tilde{x}_t = [f(\tilde{x}_t, t) - g(t)^2\nabla \ln \tilde{p}_t(\tilde{x}_t)]dt + g(t)\,d\tilde{w}_t, \ t \in [0, T], \tag{2.4}$$

where $\tilde{w}_t$ is a backward Brownian Motion Anderson, 1982. By carefully choosing f and g, we can expect that $\tilde{p}_T$ is approximately equal to some prior distribution $\tilde{q}_T$ (e.g., a centered Gaussian) which we can accurately sample from. Then we hope that starting with some $\tilde{y}_T \sim p_{\text{prior}} = \tilde{q}_T \approx \tilde{p}_T$ and running the reverse-time process, we will get a good sample $\tilde{y}_0 \sim \tilde{q}_0 \approx p_{\text{data}}$.

The case where $f \equiv 0$ and $g \equiv 1$ recovers the simple case of convolving with a Gaussian as used in §2.2.2; note, however that the reverse-time SDE differs from Langevin diffusion in having a larger (and time-varying) drift relative to the diffusion. Song, Sohl-Dickstein, et al., 2020 highlight the following two special cases. We will focus on DDPM while noting that our analysis applies more generically.

SMLD **Score-matching Langevin diffusion**: $f \equiv 0$. In this case, $\widetilde{p}_t = \widetilde{p}_0 * \varphi_{\int_0^t g(s)^2\,ds}$, so Song, Sohl-Dickstein, et al., 2020 call this a variance-exploding (VE) SDE. As is common for annealing-based algorithms, Song and Ermon, 2019; Song, Sohl-Dickstein, et al., 2020 suggest choosing an exponential schedule, so that $g(t) = ab^t$ for constants a, b. We take $p_{\text{prior}} = N(0, \int_0^T g(s)^2\,ds \cdot I_d)$.

DDPM **Denoising diffusion probabilistic modeling**: $f(x,t) = -\frac{1}{2}g(t)^2 x$. This is an Ornstein-Uhlenbeck process with time rescaling, $\widetilde{p}_t = M_{e^{-\frac{1}{2}\int_0^t g(s)^2\,ds}\sharp}\widetilde{p}_0 * \varphi_{1-e^{-\int_0^t g(s)^2\,ds}}$, where $M_\alpha(x) = \alpha x$. Song, Sohl-Dickstein, et al., 2020 call this a variance-preserving (VP) SDE, as the variance converges towards I_d. Because it displays exponential convergence towards $N(0, I_d)$, it can be run for a smaller amount of normalized time $\int_0^t g(s)^2\,ds$. Song, Sohl-Dickstein, et al., 2020 suggest the choice $g(t) = \sqrt{b + \alpha t}$. We take $p_{\text{prior}} = N(0, (1 - e^{-\int_0^t g(s)^2\,ds}) I_d) \approx N(0, I_d)$.

To obtain an algorithm, we consider the following discretization and approximation of (2.4); note that in all cases of interest the integrals can be analytically evaluated. We reverse time so that t corresponds to $T - t$ of the forward process. As we are free to rescale time in the SDE, we assume without loss of generality that the step sizes are constant. The predictor

step is

$$z_{(k+1)h} = z_{kh} - \int_{kh}^{(k+1)h} \left[f(z_{kh}, T-t) - g(T-t)^2 \cdot s(z_{kh}, T-kh) \right] dt + \int_{kh}^{(k+1)h} g(T-t)\, dw_t, \quad \text{(P)}$$

where $\int_{kh}^{(k+1)h} g(T-t)\, dw_t$ is distributed as $N(0, \int_{kh}^{(k+1)h} g(T-t)^2\, dt \cdot I_d)$. Following Song, Sohl-Dickstein, et al., 2020, we call these predictor steps as the samples aim to track the distributions $\widetilde{p}_{T-kh}$. Note that we flip the time. For simplicity of presentation, we consider the case $g \equiv 1$. We note that although the choice of the schedule does matter in practice, what really matters in our theoretical analysis is the integral $\int_0^t g(s)^2 ds$. This means that different choices of g are related by only a rescaling of time, i.e., for different g and $\tilde{g}$, we can always choose total times T and $\tilde{T}$, such that $\int_0^T g(s)^2 ds = \int_0^{\tilde{T}} \tilde{g}(s)^2 ds$. While it seems that choosing large $g(t)$ could reduce the total time T, in our analysis (e.g., Lemma 2.7.15) we need the time step-size h to be $O(1/g(T)^2)$ and hence the total computational cost, which is roughly $O(T/h)$, does not change significantly.

Theorem 2.3.1 (Predictor with L^2-accurate score estimate, DDPM)**.** *Let $p_{\text{data}} : \mathbb{R}^d \to \mathbb{R}$ be a probability density satisfying Assumption 1 with $M_2 = O(d)$, and let $\widetilde{p}_t$ be the distribution resulting from evolving the forward SDE according to DDPM with $g \equiv 1$. Suppose furthermore that $\nabla \ln \widetilde{p}_t$ is L-Lipschitz for every $t \geqslant 0$, and that each $s(\cdot, t)$ satisfies Assumption 3. Then if*

$$\varepsilon = O\left(\frac{\varepsilon_{\text{TV}}^4}{(C_{\text{LS}} + d) C_{\text{LS}}^{5/2} (L \vee L_s)^2 (\ln(C_{\text{LS}} d) \vee C_{\text{LS}} \ln(1/\varepsilon_{\text{TV}}^2))} \right),$$

running (P) *starting from p_{prior} for time $T = \Theta\left(\ln(C_{\text{LS}} d) \vee C_{\text{LS}} \ln\left(\frac{1}{\varepsilon_{\text{TV}}} \right) \right)$ and step size $h = \Theta\left(\frac{\varepsilon_{\text{TV}}^2}{C_{\text{LS}}(C_{\text{LS}}+d)(L \vee L_s)^2} \right)$ results in a distribution q_T so that $\text{TV}(q_T, p_{\text{data}}) \leqslant \varepsilon_{\text{TV}}$.*

A more precise statement of the Theorem can be found in the Appendix. Although we state our theorem for DDPM, we describe in Appendix 2.7 how it can be adapted

to other SDE's like SMLD and the sub-VP SDE; the primary SDE-dependent bound we need is a bound on $\nabla \ln \frac{\widetilde{p}_t}{\widetilde{p}_{t+h}}$. Because the predictor is tracking a changing distribution p_t, we incur more error terms and worse dependence on parameters (C_{LS}, L) than in LMC (Theorem 2.2.1). Motivated by this, we intersperse the predictor steps with LMC steps—called *corrector* steps in this context—to give additional time for the process to mix, resulting in improved dependence on parameters.

Algorithm 2 Predictor-corrector method with estimated score Song, Sohl-Dickstein, et al., 2020

INPUT: Time T, predictor step size h; number of corrector steps N_m per predictor step, corrector step sizes h_m
Draw $z_0 \sim p_{\mathrm{prior}}$ from the prior distribution.
for m from 1 to T/h **do**
(Predictor) Take a step of (P) to obtain z_{mh} from $z_{(m-1)h}$, with f, g as in SMLD or DDPM.

(Corrector) Starting from $z_{mh,0} := z_{mh}$, run (LMC-SE) with $s(z, T - mh)$ and step size h_m for N steps, and let $z_{mh} \leftarrow z_{mh,N}$.
end for
OUTPUT: Return z_T, approximate sample from p_{data}.

Theorem 2.3.2 (Predictor-corrector with L^2-accurate score estimate)**.** *Keep the setup of Theorem 2.3.1. Then for* $\varepsilon_{\mathrm{TV}}^3 = O\left(\frac{1}{(1+L_s/L)^2(1+C_{\mathrm{LS}}/d)(\ln(C_{\mathrm{LS}}d)\vee C_{\mathrm{LS}})}\right)$*, if*

$$\varepsilon = O\left(\frac{\varepsilon_{\mathrm{TV}}^4}{dL^2 C_{\mathrm{LS}}^{5/2} \ln(1/\varepsilon_\chi^2)}\right), \tag{2.5}$$

then Algorithm 2 with appropriate choices of $T = \Theta\left(\ln(C_{\mathrm{LS}}d) \vee C_{\mathrm{LS}} \log\left(\frac{1}{\varepsilon_{\mathrm{TV}}}\right)\right)$*,* N_m*, corrector step sizes* h_m *and predictor step size* h*, produces a sample from a distribution* q_T *such that* $\mathrm{TV}(q_T, p_{\mathrm{data}}) < \varepsilon_{\mathrm{TV}}$.

The assumption on $\varepsilon_{\mathrm{TV}}$ is for convenience in stating our bound. In comparison to using the predictor step alone (Theorem 2.3.1), note that in the bound on ε, we obtain the improved rate of the corrector step as in Theorem 2.2.1; this is because the predictor step only needs to track the actual distribution in χ^2-divergence with error $O(1)$, and the final corrector steps are responsible for decreasing the error to $\varepsilon_{\mathrm{TV}}$. In comparison to the Annealed Langevin

sampler (Algorithm 1, Theorem 2.2.2), which can be viewed as using the corrector step alone, adding a predictor step provides a better warm start for the distribution at the next smaller noise level, resulting in better dependence on parameters. Thus the predictor-corrector algorithm combines the strengths of the predictor and corrector steps. For real-world data, it can be challenging to estimate TV-distance between distributions given only samples, and hence difficult to check consistency with empirical observations. However, our claim that using a corrector can improve the convergence rate of DDPM/SMLD is consistent with the simulation results in Section 4.2 of Song, Sohl-Dickstein, et al., 2020.

2.4 Theoretical framework and proof sketches

The main idea of our analysis framework is to convert a L^2 error guarantee to a L^∞ error guarantee by excluding a bad set, formalized in the following theorem.

Theorem 2.4.1. *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 $\overline{Z}_n \sim \bar{q}_n$ 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 $0 \leqslant k \leqslant n-1$, then $Z_n = \overline{Z}_n$. (For $n = 0$, this says $Z_0 = \overline{Z}_0$.)*
2. *$\chi^2(\bar{q}_n || p_n) \leqslant D_n^2$.*
3. *$\mathbb{P}(X_n \in B_n) \leqslant \delta_n$.*

Then the following hold.

$$\mathrm{TV}(q_n, \bar{q}_n) \leqslant \sum_{k=0}^{n-1} (D_k^2 + 1)^{1/2} \delta_k^{1/2} \qquad \mathrm{TV}(p_n, q_n) \leqslant D_n + \sum_{k=0}^{n-1} (D_k^2 + 1)^{1/2} \delta_k^{1/2} \tag{2.6}$$

For our setting, we will take the "bad sets" B_n to be the set of x where $\|s_\theta(x) - \nabla \ln p\|$ is large, q_n to be the discretized process with estimated score, and $\bar{q}_n$ to be the discretized process with estimated score except in B_n where the error is large. Because $\bar{q}_n$ uses an L^∞-accurate score estimate, we can use existing techniques for analyzing Langevin Monte Carlo (Chewi et al., 2021; Erdogdu et al., 2021; Vempala & Wibisono, 2019) to bound $\chi^2(\bar{q}_n || p_n)$.

Proof. First note that if some $Z_k \in B_k$ for $0 \leqslant k \leqslant n-1$, then for the smallest such k, we have $\overline{Z}_k = Z_k \in B_k$; the same is true if $\overline{Z}_k \in B_k$ for some $0 \leqslant k \leqslant n-1$. We then bound using condition 1 and Cauchy-Schwarz:

$$\begin{aligned}\mathbb{P}\left(Z_n \neq \overline{Z}_n\right) &\leqslant \mathbb{P}\left(\bigcup_{k=0}^{n-1}\{Z_k \in B_k\}\right) = \mathbb{P}\left(\bigcup_{k=0}^{n-1}\{\overline{Z}_k \in B_k\}\right) \\ &\leqslant \sum_{k=0}^{n-1}\mathbb{P}\left(\overline{Z}_k \in B_k\right) = \sum_{k=0}^{n-1}\mathbb{E}_{\overline{q}_k}1_{B_k} \\ &\leqslant \sum_{k=0}^{n-1}\left(\mathbb{E}_{p_k}\left(\frac{\overline{q}_k}{p_k}\right)^2\right)^{1/2}\left(\mathbb{E}_{p_k}1_{B_k}\right)^{1/2} = \sum_{k=0}^{n-1}(D_k^2+1)^{1/2}\delta_k^{1/2}.\end{aligned}$$

The second inequality then follows from the triangle inequality and Cauchy-Schwarz:

$$\begin{aligned}\mathrm{TV}(p_n, q_n) &\leqslant \mathrm{TV}(p_n, \overline{q}_n) + \mathrm{TV}(\overline{q}_n, q_n) \\ &\leqslant \sqrt{\chi^2(\overline{q}_n||p_n)} + \mathrm{TV}(\overline{q}_n, q_n) \leqslant D_n + \sum_{k=0}^{n-1}(D_k^2+1)^{1/2}\delta_k^{1/2}.\end{aligned}$$ □

It now remains to give χ^2 convergence bounds under L^∞-accurate score estimate. The following theorem may be of independent interest.

Theorem 2.4.2 (LMC under L^∞ bound on gradient error)**.** *Let $p : \mathbb{R}^d \to \mathbb{R}$ be a probability density satisfying Assumption 1(1, 2) and $s : \mathbb{R}^d \to \mathbb{R}^d$ be a score estimate s with error bounded in L^∞: for some $\varepsilon_1 \leqslant \sqrt{\frac{1}{48C_{\mathrm{LS}}}}$,*

$$\|\nabla \ln p - s\|_\infty = \max_{x\in\mathbb{R}^d}\|\nabla \ln p(x) - s(x)\|] \leqslant \varepsilon_1.$$

Let $N \in \mathbb{N}_0$ and $0 < h \leqslant \frac{1}{4392dC_{\mathrm{LS}}L^2}$, and assume $L \geqslant 1$. Let q_{nh} denote the nth iterate of LMC with step size h score estimate s. Then

$$\chi^2(q_{(k+1)h}||p) \leqslant \exp\left(-\frac{h}{4C_{\mathrm{LS}}}\right)\chi^2(q_{kh}||p) + 170dL^2h^2 + 5\varepsilon_1^2 h$$

and

$$\chi^2(q_{Nh}||p) \leqslant \exp\left(-\frac{Nh}{4C_{\mathrm{LS}}}\right)\chi^2(q_0||p) + 680dL^2hC_{\mathrm{LS}} + 20\varepsilon_1^2C_{\mathrm{LS}}$$

$$\leqslant \exp\left(-\frac{Nh}{4C_{\mathrm{LS}}}\right)\chi^2(q_0||p) + 1.$$

Following (Chewi et al., 2021), we prove this by first defining a continuous-time interpolation q_t of the discrete process, and then deriving a differential inequality for $\chi^2(q_t||p)$ using the log-Sobolev inequality for p. Compared to (Chewi et al., 2021), we incur an extra error term arising from the inaccurate gradient.

This allows us to sketch the proof of Theorem 2.2.1; a complete proof is in Section 2.6.

Proof sketch of Theorem 2.2.1. We first define the bad set where the error in the score estimate is large,

$$B := \{\|\nabla \ln p(x) - s(x)\| > \varepsilon_1\}$$

for some ε_1 to be chosen. Then by Chebyshev's inequality, $P(B) \leqslant \left(\frac{\varepsilon}{\varepsilon_1}\right)^2 =: \delta$. Let $\overline{q}_{nh}$ be the discretized process, but where the score estimate is set to be equal to $\nabla \ln p$ on B; note it agrees with q_{nh} as long as it has not hit B. Because $\overline{q}_{nh}$ uses a score estimate that has L^∞-error ε_1, Theorem 2.4.2 gives a bound for $\chi^2(\overline{q}_{Nh}||p)$. Then Theorem 2.4.1 gives

$$\mathrm{TV}(q_{nh}, \overline{q}_{nh}) \leqslant \sum_{k=0}^{n-1}(\chi^2(\overline{q}_{kh}||p)+1)^{1/2}P(B)^{1/2} \leqslant \sum_{k=0}^{n-1}\left(\exp\left(-\frac{kh}{8C_{\mathrm{LS}}}\right)\chi^2(q_0||p)^{1/2}+1\right)\delta^{1/2}$$

The theorem then follows from choosing parameters so that $\chi^2(\overline{q}_T||p) \leqslant \varepsilon_\chi^2$ and $\mathrm{TV}(q_T, \overline{q}_T) \leqslant \varepsilon_{\mathrm{TV}}$. □

We remark that the main inefficiency in the proof comes from the use of Chebyshev's inequality, and a L^p bound on the error for $p > 2$ will improve the bound.

Proof sketch of Theorem 2.2.2. Choosing the sequence $\sigma_1 < \cdots < \sigma_M$ to be geometric with ratio $1 + \frac{1}{\sqrt{d}}$ ensures that the χ^2-divergence between successive distributions $p_{\sigma_m^2}$ is $O(1)$. Then, choosing $\sigma_M^2 = \Omega(C_{\mathrm{LS}}d)$ ensures we have a warm start for the highest noise level:

$\chi^2(p_{\text{prior}}||p_{\sigma_M^2}) = O(1)$. This uses $O\left(\sqrt{d}\log\left(\frac{dC_{\text{LS}}}{\sigma_{\min}^2}\right)\right)$ noise levels. Chebyshev's inequality can be used to show that the distribution of the final sample $x^{(m)}$ for $p_{\sigma_m^2}$ is $O(\varepsilon_{\text{TV}}/M)$ close to a distribution that is $O(M/\varepsilon_{\text{TV}})$ in χ^2-divergence from $p_{\sigma_{m+1}^2}$. This gives the warm start parameter $K_\chi = (M/\varepsilon_{\text{TV}})^{1/2}$; substituting into Theorem 2.2.1 then gives the required bound for ε. Note that the TV errors accrued from each level add to $O(\varepsilon_{\text{TV}})$. □

To analyze the predictor-based algorithms, we also first prove convergence bounds under L^∞-accurate score estimate.

Theorem 2.4.3 (Predictor steps under L^∞ bound on score estimate, DDPM)**.** *Let $p : \mathbb{R}^d \to \mathbb{R}$ be a probability density satisfying Assumption 1 and $s(\cdot, t) : \mathbb{R}^d \to \mathbb{R}^d$ be a score estimate s with error bounded in L^∞ for each $t \in [0, T]$:*

$$\|\nabla \ln p - s(\cdot, t)\|_\infty = \max_{x\in\mathbb{R}^d} \|\nabla \ln \widetilde{p}_t(x) - s(x, t)\|] \leqslant \varepsilon_1.$$

Consider DDPM with $g \equiv 1$, $T \geqslant 1 \vee \ln(C_{\text{LS}}d)$, and $h = O\left(\frac{1}{C_{\text{LS}}(d+C_{\text{LS}})(L\vee L_s)^2}\right)$. (Recall that p_{kh} and q_{kh} are the k-th iterate of LMC with step size h and true/estimated score respectively.) Then

$$\chi^2(q_{(k+1)h}||p_{(k+1)h}) \leqslant \chi^2(q_{kh}||p_{kh})e^{\left(-\frac{1}{8C_{\text{LS}}}+8\varepsilon_1^2\right)h} + O(\varepsilon_1^2 h + (L_s^2 + L^2 d)h^2)$$

and if $\varepsilon_1 < \frac{1}{128C_{\text{LS}}}$,

$$\chi^2(q_{Nh}||p_{Nh}) \leqslant e^{-\frac{Nh}{16C_{\text{LS}}}}\chi^2(q_0||p_0) + O\left(C_{\text{LS}}\left(\varepsilon_1^2 + (L_s^2 + L^2 d)h\right)\right).$$

Moreover, for $q_0 = p_{\text{prior}}$, $\chi^2(q_0||p_0) \leqslant e^{-T/2}C_{\text{LS}}d$.

We give a more precise statement in Section 2.7. Note that unlike the case for LMC as in Theorem 2.4.2, the base density p_t is also evolving in time, which produces additional error terms and necessitates a more involved analysis. The additional error terms can be bounded using the Donsker-Varadhan variational principle, concentration for distributions satisfying LSI, and error bounds between p_t and p_{t+h} for small h.

Here, we only state the result about DDPM, which has better bounds than SMLD (when $g \equiv 1$) because both the forward and backwards processes exhibit better mixing properties:

the warm start improves exponentially rather than inversely with T, and the log-Sobolev constant is uniformly bounded by that of p_{data} rather than increasing. However, the analysis in Section 2.7 can be directly applied to SMLD and other models as well. We also note there is a sense in which DDPM and SMLD are equivalent under a rescaling in time and space (see discussion in Section 2.7.2).

Note that the choice of h is necessary for exponential decay of error; as if h is not small enough, we would get an exponential growing instead of decaying factor in the one-step error (See Section 2.7 for details). Such an h may however still be a suitable choice when used in conjunction with a corrector step. Moreover, as $\varepsilon_1 \to 0$, with appropriate choice of T and h, q_{Nh} and p_{Nh} can be made arbitrarily close.

Theorem 2.3.1 now follows from the L^∞ result (Theorem 2.4.3) in the same way that Theorem 2.2.1 follows from Theorem 2.4.2.

To prove Theorem 2.3.2, it suffices to run the corrector steps only at the lowest noise level, that is, set $N_m = 0$ for $1 \leqslant m < T/h$, although we note that interleaving the predictor and corrector steps does empirically help with mixing. The proof follows from using the predictor and the corrector theorems in series: first apply Theorem 2.3.1 with $\varepsilon_\chi = O(1)$ to show that the predictor results a warm start p_{data}, then use Theorem 2.2.1 to show the corrector reduces the error to the desired ε_{TV}.

2.5 Computations

We start the proofs by collecting some preliminary results. In the following, we will consider the SDE

$$dx_t = f(x,t)\,dt + G(t)\,dw_t \tag{2.7}$$

and the interpolation of the discretization of an approximation

$$dz_t = \widehat{f}(z_{t_-},t)\,dt + G(t)\,dw_t \tag{2.8}$$

when $t \geqslant t_-$. Let P_t and Q_t denote the law of x_t and z_t, respectively. We will take $t_- = kh$ and $t \in [kh,(k+1)h)$. We will assume that $f, \widehat{f}, G$ are continuous and the functions $f(\cdot,t)$, $\widehat{f}(\cdot,t)$ are uniformly Lipschitz for each $t \in [kh,(k+1)h]$.

In this section, we will make some computations that will be used in both Sections 2.6 and 2.7. First, we derive how the density evolves in time.

Lemma 2.5.1. *Let Q_t denote the law of the interpolated process* (2.8)*. Then*

$$\frac{\partial q_t(z)}{\partial t} = \nabla \cdot \left[-q_t(z)\mathbb{E}\left[\widehat{f}(z_{t\_}, t)|z_t = z\right] + \frac{G(t)G(t)^\top}{2}\nabla q_t(z)\right].$$

Proof. Let $q_{t|t\_}$ denote the distribution of z_t conditioned on $z_{t\_}$. Then the Fokker-Planck equation gives

$$\frac{\partial q_{t|t\_}(z|z_{t\_})}{\partial t} = -\nabla q_{t|t\_}(z|z_{t\_}) \cdot [\widehat{f}(z_{t\_}, t)] + \frac{G(t)G(t)^\top}{2}\Delta q_{t|t\_}(z|z_{t\_})$$

Taking expectation with respect to $z_{t\_}$ we get

$$\begin{aligned}\frac{\partial q_t(z)}{\partial t} &= \nabla \cdot \int -q_{t|t\_}(z)\widehat{f}(y,t)q_{t\_}(y)dy + \nabla \cdot \left[\frac{G(t)G(t)^\top}{2}\nabla \int q_{t|t\_}(z|y)q_{t\_}(y)dy\right] \\ &= \nabla \cdot q_t(z) \int \left[-\widehat{f}(y,t)q_{k|t}(y|z)dy + \frac{G(t)G(t)^\top}{2}\nabla \int q_{t\_|t}(y|z)q_t(z)dy\right].\end{aligned}$$

Note that for fixed z, $\int q_{t\_|t}(y|z)dy = 1$. Hence

$$\frac{\partial q_t(z)}{\partial t} = \nabla \cdot \left[-q_t(z)\mathbb{E}\left[\widehat{f}(z_{t\_}, t)|z_t = z\right] + \frac{G(t)G(t)^\top}{2}\nabla q_t(z)\right].$$

□

We now use Lemma 2.5.1 to compute how the χ^2-divergence between the approximate and exact densities changes. The following generalizes the calculation of (Erdogdu et al., 2021) in the case where x_t is a non-stationary stochastic process. For simplicity of notation, from now on, we wil consider the case $G(t)$ being a scalar.

Lemma 2.5.2. *Let P_t and Q_t be the laws of* (2.7) *and* (2.8) *for $G(t) = g(t)I_d$. Then*

$$\frac{\partial}{\partial t}\chi^2(q_t||p_t) = -g(t)^2\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 2\mathbb{E}\left[\left\langle \widehat{f}(z_{t\_}, t) - f(z_t, t), \nabla\frac{q_t(z_t)}{p_t(z_t)}\right\rangle\right].$$

Proof. The Fokker-Planck equation gives

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

We have

$$\frac{d}{dt}\chi^2(q_t||p_t) = \frac{d}{dt}\int \frac{q_t(x)^2}{p_t(x)}dx = \int \left[2\frac{\partial q_t(x)}{\partial t}\frac{q_t(x)}{p_t(x)} - \frac{\partial p_t(x)}{\partial t}\frac{q_t(x)^2}{p_t(x)^2}\right] dx.$$

For the first term, by Lemma 2.5.1,

$$\begin{aligned}
2\int \frac{\partial q_t(x)}{\partial t}\frac{q_t(x)}{p_t(x)}dx &= 2\int \nabla\cdot\left[-q_t(x)\mathbb{E}\left[\widehat{f}(z_0,t)|z_t=x\right] + \frac{g(t)^2}{2}\nabla q_t(x)\right]\cdot\frac{q_t(x)}{p_t(x)}dx \\
&= 2\int q_t(x)\left\langle \mathbb{E}\left[\widehat{f}(z_0,t)|z_t=x\right], \nabla\frac{q_t(x)}{p_t(x)}\right\rangle dx \\
&\quad - g(t)^2\int\left\langle \nabla q_t(x), \nabla\frac{q_t(x)}{p_t(x)}\right\rangle dx. \qquad (2.9)
\end{aligned}$$

For the second term, using integration by parts,

$$\begin{aligned}
-\int \frac{\partial p_t(x)}{\partial t}\frac{q_t(x)^2}{p_t(x)^2}dx &= \int \nabla\cdot\left[f(x,t)p_t(x) - \frac{g(t)^2}{2}\nabla p_t(x)\right]\cdot\frac{q_t(x)^2}{p_t(x)^2}dx \\
&= \int -f(x,t)p_t(x)\nabla\frac{q_t(x)^2}{p_t(x)^2} + \frac{g(t)^2}{2}\left\langle \nabla p_t(x), \nabla\frac{q_t(x)^2}{p_t(x)^2}\right\rangle dx \\
&= -2\int q_t(x)\left\langle f(x,t), \nabla\frac{q_t(x)}{p_t(x)}\right\rangle dx \\
&\quad + g(t)^2\int\frac{q_t(x)}{p_t(x)}\left\langle \nabla p_t(x), \nabla\frac{q_t(x)}{p_t(x)}\right\rangle dx. \qquad (2.10)
\end{aligned}$$

Note that

$$\int\left\langle \nabla q_t(x), \nabla\frac{q_t(x)}{p_t(x)}\right\rangle - \frac{q_t(x)}{p_t(x)}\left\langle \nabla p_t(x), \nabla\frac{q_t(x)}{p_t(x)}\right\rangle = \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right).$$

Combining (2.9) and (2.10),

$$\begin{aligned}
\frac{d}{dt}\chi^2(q_t||p_t) &= -g(t)^2\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 2\int q_t(x)\left\langle \mathbb{E}\left[\widehat{f}(z_{t_},t) - f(x,t)|z_t=x\right], \nabla\frac{q_t(x)}{p_t(x)}\right\rangle dx \\
&= -g(t)^2\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 2\mathbb{E}\left[\left\langle \widehat{f}(z_{t_},t) - f(z_t,t), \nabla\frac{q_t(z_t)}{p_t(z_t)}\right\rangle\right].
\end{aligned}$$

□

Finally, we will make good use of the following lemma to bound the second term in Lemma 2.5.2.

Lemma 2.5.3 (cf. (Erdogdu et al., 2021, Lemma 1)). *Let $\phi_t(x) = \frac{q_t(x)}{p_t(x)}$ and $\psi_t(x) = \phi_t(x)/\mathbb{E}_{p_t}\phi_t^2$. For any c and any $\mathbb{R}^d$-valued random variable u, we have*

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

Proof. Note that $\mathbb{E}\psi_t(z_t) = 1$ and the normalizing factor is $\mathbb{E}_{p_t}\phi_t^2 = \chi^2(q_t||p_t) + 1$. By Young's inequality,

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

□

2.6 Analysis for LMC

Let p be the probability density we wish to sample from. Suppose that we have an estimate s of the score $\nabla \ln p$. Our main theorem says that if the L^2 error $\mathbb{E}_p \|\nabla \ln p - s\|^2$ is small enough, then running LMC with s for an *appropriate* time results in a density that is close in *TV distance* to a density that is close in χ^2*-divergence* to p. The following is a more precise version of Theorem 2.2.1.

Theorem 2.6.1 (LMC with L^2-accurate score estimate). *Let $p : \mathbb{R}^d \to \mathbb{R}$ be a probability density satisfying Assumption 1 with $L \geqslant 1$ and $s : \mathbb{R}^d \to \mathbb{R}^d$ be a score estimate satisfying Assumption 3(2). Consider the accuracy requirement in* TV *and χ^2: $0 < \varepsilon_{\mathrm{TV}} < 1$, $0 < \varepsilon_\chi < 1$, and suppose furthermore the starting distribution satisfies $\chi^2(p_0||p) \leqslant K_\chi^2$. Then if*

$$\varepsilon \leqslant \frac{\varepsilon_{\mathrm{TV}}\varepsilon_\chi^3}{174080\sqrt{5}dL^2C_{\mathrm{LS}}^{5/2}(C_T \ln(2K_\chi/\varepsilon_\chi^2) \vee 2K_\chi)},$$

then running (LMC-SE) *with score estimate s and step size $h = \frac{\varepsilon_\chi^2}{2720dL^2C_{\mathrm{LS}}}$ for any time $T \in [T_{\min}, C_T T_{\min}]$, where $T_{\min} = 4C_{\mathrm{LS}} \ln\left(\frac{2K_\chi}{\varepsilon_\chi^2}\right)$, results in a distribution p_T such that p_T is $\varepsilon_{\mathrm{TV}}$-far*

in TV distance from a distribution $\overline{p}_T$, where $\overline{p}_T$ satisfies $\chi^2(\overline{p}_T||p) \leqslant \varepsilon_\chi^2$. In particular, taking $\varepsilon_\chi = \varepsilon_{\mathrm{TV}}$, we have the error guarantee that $\mathrm{TV}(p_T, p) = 2\varepsilon_{\mathrm{TV}}$.

The main difficulty is that the stationary distribution of LMC using the score estimate may be arbitrarily far from p, even if the L^2 error of the score estimate is bounded. (See Section 2.8.) Thus, a long-time convergence result does not hold, and an upper bound on T is required, as in the theorem statement.

We instead proceed by showing that *conditioned on not hitting a bad set*, if we run LMC using s, the χ^2-divergence to the stationary distribution will decrease. This means that the closeness of the overall distribution (in TV distance, say) will decrease in the short term, despite it will increase in the long term, as the probability of hitting the bad set increases. This does not contradict the fact that the stationary distribution is different from p. By running for a moderate amount of time (just enough for mixing), we can ensure that the probability of hitting the bad set is small, so that the resulting distribution is close to p. Note that we state the theorem with a C_T parameter to allow a range of times that we can run LMC for.

More precisely, we prove Theorem 2.6.1 in two steps.

LMC under L^∞ gradient error (Section 2.6.1, Theorem 2.4.2). First, consider a simpler problem: proving a bound for χ^2 divergence for LMC with score estimate s, when $\|s - \nabla \ln p\|$ is bounded everywhere, not just on average. For this, we follow the argument in (Chewi et al., 2021) for showing convergence of LMC in Rényi divergence; this also gives a bound in χ^2-divergence. We define an interpolation of the discrete process and derive a upper bound for the derivative of Rényi divergence, $\partial_t \mathcal{R}_q(q_t||p)$, using the log-Sobolev inequality for p. In the original proof, the error comes from the discretization error; here we have an additional error term coming from an inaccurate gradient, which is bounded by assumption. Note that a L^2 bound on $\nabla f - s$ is insufficient to give an upper bound, as we need to bound $\mathbb{E}_{q_t\psi_t}[\|\nabla f - s\|^2]$ for a different measure $q_t\psi_t$ that we do not have good control over. An L^∞ bound works regardless of the measure.

Defining a bad set and bounding the hitting time (Section 2.6.2). The idea is now to reduce to the case of L^∞ error by defining the "bad set" B to be the set where $\|s - \nabla f\| \geqslant \varepsilon_1$, where $\varepsilon \ll \varepsilon_1 \ll 1$. This set has small measure by Chebyshev's inequality. Away from the bad set, Theorem 2.4.2 applies; it then suffices to bound the probability of hitting B. Technically, we define a coupling with a hypothetical process where the L^∞ error is always bounded, and note that the processes disagree exactly when it hits B; this is the source of the TV error.

We consider the probability of being in B at times $0, h, 2h, \ldots$. we note that Theorem 2.6.1 bounds the χ^2-divergence of this hypothetical process X_t at time t to p. If the distribution were actually p, then the probability $X_t \in B'$ is exactly $p(B')$; we expect the probability to be small even if the distribution is close to p. Indeed, by Cauchy-Schwarz, we can bound the probability $X \in B$ in terms of $P(B)$ and $\chi^2(q_t||p)$; this bound is given in Theorem 3.7.1. Note that the eventual bound depends on $\chi^2(q_t||p)$, so we have to assume a warm start, that is, a reasonable bound on $\chi^2(q_0||p)$.

2.6.1 LMC under L^∞ gradient error

The following gives a long-time convergence bound for LMC with inaccurate gradient, with error bounded in L^∞; this may be of independent interest.

Theorem 2.4.2 (LMC under L^∞ bound on gradient error)**.** *Let $p : \mathbb{R}^d \to \mathbb{R}$ be a probability density satisfying Assumption 1(1, 2) and $s : \mathbb{R}^d \to \mathbb{R}^d$ be a score estimate s with error bounded in L^∞: for some $\varepsilon_1 \leqslant \sqrt{\frac{1}{48C_{\mathrm{LS}}}}$,*

$$\|\nabla \ln p - s\|_\infty = \max_{x \in \mathbb{R}^d} \|\nabla \ln p(x) - s(x)\|] \leqslant \varepsilon_1.$$

Let $N \in \mathbb{N}_0$ and $0 < h \leqslant \frac{1}{4392dC_{\mathrm{LS}}L^2}$, and assume $L \geqslant 1$. Let q_{nh} denote the nth iterate of LMC with step size h score estimate s. Then

$$\chi^2(q_{(k+1)h}||p) \leqslant \exp\left(-\frac{h}{4C_{\mathrm{LS}}}\right)\chi^2(q_{kh}||p) + 170dL^2h^2 + 5\varepsilon_1^2 h$$

and

$$\chi^2(q_{Nh}||p) \leqslant \exp\left(-\frac{Nh}{4C_{\mathrm{LS}}}\right)\chi^2(q_0||p) + 680dL^2hC_{\mathrm{LS}} + 20\varepsilon_1^2C_{\mathrm{LS}}$$
$$\leqslant \exp\left(-\frac{Nh}{4C_{\mathrm{LS}}}\right)\chi^2(q_0||p) + 1.$$

Following (Chewi et al., 2021), convergence in Rényi divergence can also be derived; we only consider χ^2-divergence because we will need a warm start in χ^2-divergence for our application. Note that by letting $N \to \infty$ and $h \to 0$, we obtain the following.

Corollary 2.6.2. *Keep the assumptions in Theorem 2.4.2. The stationary distribution q of Langevin diffusion with score estimate s satisfies*

$$\chi^2(q||p) \leqslant 20C_{\mathrm{LS}}\varepsilon_1^2.$$

Proof of Theorem 2.4.2. We follow the proof of (Chewi et al., 2021, Theorem 4), except that we work with the χ^2 divergence directly, rather than the Rényi divergence, and have an extra term from the inaccurate gradient (2.17). Given $t \geqslant 0$, let $t_- = h\left\lfloor \frac{t}{h} \right\rfloor$. Define the interpolated process by

$$dz_t = s(z_{t_-})\,dt + \sqrt{2}\,dw_t, \tag{2.11}$$

and let q_t denote the distribution of X_t at time t, when $X_0 \sim q_0$.

By Lemma 2.5.2,

$$\frac{\partial}{\partial t}\chi^2(q_t||p) = -2\mathscr{E}_p\left(\frac{q_t}{p}\right) + 2\mathbb{E}\left[\left\langle s(z_{t_-}) - \nabla \ln p(z_t), \nabla\frac{q_t(z_t)}{p(z_t)}\right\rangle\right]. \tag{2.12}$$

By the proof of Theorem 4 in (Chewi et al., 2021),

$$\left\|\nabla \ln p(x_t) - \nabla \ln p(x_{t_-})\right\|^2 \leqslant 9L^2(t-t_-)^2\left\|\nabla \ln p(x_t)\right\|^2 + 6L^2\left\|B_t - B_{t_-}\right\|^2.$$

Then

$$\left\|s(z_{t_-}) - \nabla \ln p(z_t)\right\|^2 \leqslant 2\left\|\nabla \ln p(z_{t_-}) - \nabla \ln p(z_t)\right\|^2 + 2\left\|s(z_{t_-}) - \nabla \ln p(z_{t_-})\right\|^2$$
$$\leqslant 18L^2(t-t_-)^2\left\|\nabla \ln p(z_t)\right\|^2 + 12L^2\left\|B_t - B_{t_-}\right\|^2 + 2\varepsilon_1^2. \tag{2.13}$$

Let $\phi_t := q_t/p$ and $\psi_t := \frac{\phi_t}{\mathbb{E}_p(\phi_t^2)}$. By Lemma 2.5.3,

$$2\mathbb{E}\left[\left\langle s(z_{t\_}) - \nabla \ln p(z_t), \nabla \frac{q_t(z_t)}{p(z_t)} \right\rangle\right] \leqslant 2\mathbb{E}_p\phi_t^2 \cdot \mathbb{E}\left[\left\|s(z_{t\_}) - \nabla \ln p(z_t)\right\|^2 \psi_t(z_t)\right] + \frac{1}{2}\mathscr{E}_p\left(\frac{q_t}{p}\right)$$

$$\leqslant A_1 + A_2 + A_3 + \frac{1}{2}\mathscr{E}_p(\phi_t) \tag{2.14}$$

where A_1, A_2, A_3 are obtained by substituting in the 3 terms in (2.13), and given in (2.15), (2.16), and (2.17). Let $V(x) = -\ln p(x)$. We consider each term in turn.

$$A_1 := 36L^2(t-t_{\_})^2\mathbb{E}_p\phi_t^2 \cdot \mathbb{E}\left[\left\|\nabla V(z_t)\right\|^2 \psi_t(z_t)\right] \tag{2.15}$$

$$\leqslant 36L^2(t-t_{\_})^2\mathbb{E}_p\phi_t^2 \cdot \left(\frac{4\mathscr{E}_p(\phi_t)}{\mathbb{E}_p\phi_t^2} + 2dL\right) \qquad \text{by (Chewi et al., 2021, Lemma 16)}$$

$$\leqslant \frac{1}{2}\mathscr{E}_p(\phi_t) + 72dL^3(t-t_{\_})^2(\chi^2(q_t||p) + 1)$$

when $h^2 \leqslant \frac{1}{288L^2}$. By (Chewi et al., 2021, p. 15)

$$A_2 := 24L^2\mathbb{E}_p\phi_t^2 \cdot \mathbb{E}\left[\left\|B_t - B_{t\_}\right\|^2 \psi_t(z_t)\right] \tag{2.16}$$

$$\leqslant 24L^2\mathbb{E}_p\phi_t^2 \cdot \left(14dL^2(t-t_{\_}) + 32hC_{\mathrm{LS}}\frac{\mathscr{E}_p(\phi_t)}{\mathbb{E}_p\phi_t^2}\right)$$

$$\leqslant 336dL^2(t-t_{\_})(\chi^2(q_t||p) + 1) + \frac{1}{2}\mathscr{E}_p(\phi_t)$$

when $h \leqslant \frac{1}{1536L^2C_{\mathrm{LS}}}$. Finally,

$$A_3 := 4\varepsilon_1^2\mathbb{E}_p\phi_t^2 = 4\varepsilon_1^2(\chi^2(q_t||p) + 1). \tag{2.17}$$

Combining (2.12), (2.14), (2.15), (2.16), and (2.17) gives

$$\frac{\partial}{\partial t}\chi^2(q_t||p) \leqslant -\frac{1}{2}\mathscr{E}_p(\phi_t) + (\chi^2(q_t||p) + 1)(72dL^3(t-t_{\_})^2 + 336dL^2(t-t_{\_}) + 4\varepsilon_1^2)$$

$$\leqslant -\frac{1}{2C_{\mathrm{LS}}}\chi^2(q_t||p) + (\chi^2(q_t||p) + 1)(72dL^3(t-t_{\_})^2 + 336L^2d(t-t_{\_}) + 4\varepsilon_1^2)$$

$$\leqslant -\frac{1}{4C_{\mathrm{LS}}}\chi^2(q_t||p) + (72dL^3(t-t_{\_})^2 + 336dL^2(t-t_{\_}) + 4\varepsilon_1^2)$$

if $h \leqslant \left(\frac{1}{12\cdot 72dL^3C_{\mathrm{LS}}}\right)^{1/2} \wedge \frac{1}{12\cdot 336dC_{\mathrm{LS}}}$ and $\varepsilon_1 \leqslant \left(\frac{1}{48C_{\mathrm{LS}}}\right)^{1/2}$. Then for $t \in [kh, (k+1)h)$,

$$\frac{\partial}{\partial t}\left(\chi^2(q_t||p)\exp\left(\frac{t-t_-}{4C_{\mathrm{LS}}}\right)\right) = \exp\left(\frac{t-t_-}{4C_{\mathrm{LS}}}\right)(72dL^3(t-t_-)^2 + 336dL^2(t-t_-) + 4\varepsilon_1^2)$$

$$\leqslant 73dL^3(t-t_-)^2 + 337dL^2(t-t_-) + 5\varepsilon_1^2.$$

Integrating over $t \in [kh, (k+1)h)$ gives

$$\chi^2(q_{(k+1)h}||p) \leqslant \exp\left(-\frac{h}{4C_{\mathrm{LS}}}\right)\chi^2(q_{kh}||p) + \frac{73}{3}dL^3h^3 + \frac{337}{2}dL^2h^2 + 5\varepsilon_1^2 h$$

$$\leqslant \exp\left(-\frac{h}{4C_{\mathrm{LS}}}\right)\chi^2(q_{kh}||p) + 170dL^2h^2 + 5\varepsilon_1^2 h$$

using $h \leqslant \frac{1}{12\sqrt{2}L}$. Unfolding the recurrence and summing the geometric series gives

$$\chi^2(q_{kh}||p) \leqslant \exp\left(-\frac{kh}{4C_{\mathrm{LS}}}\right)\chi^2(q_0||p) + 680dL^2hC_{\mathrm{LS}} + 20\varepsilon_1^2C_{\mathrm{LS}}$$

$$\leqslant \exp\left(-\frac{kh}{4C_{\mathrm{LS}}}\right)\chi^2(q_0||p) + 1$$

when $h \leqslant \frac{1}{1360dL^2C_{\mathrm{LS}}}$ and $\varepsilon_1^2 \leqslant \frac{1}{40C_{\mathrm{LS}}}$. We can check that the given condition on h and the fact that $LC_{\mathrm{LS}} \geqslant 1$ (Lemma 2.9.5) imply all the required inequalities on h. □

2.6.2 Proof of Theorem 2.6.1

Proof of Theorem 2.6.1. We first define the bad set where the error in the score estimate is large,

$$B := \{\|\nabla \ln p(x) - s(x)\| > \varepsilon_1\}$$

for some ε_1 to be chosen.

Given $t \geqslant 0$, let $t_- = h\left\lfloor \frac{t}{h} \right\rfloor$. Given a bad set B, define the interpolated process by

$$d\overline{z}_t = b(\overline{z}_{t_-})\,dt + \sqrt{2}\,dw_t, \tag{2.18}$$

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

In other words, run LMC using the score estimate as long as the point is in the good set at the previous discretization step, and otherwise use the actual gradient $\nabla \ln p$. Let $\overline{q}_t$ denote

the distribution of $\bar{z}_t$ when $\bar{z}_0 \sim q_0$; note that q_{nh} is the distribution resulting from running LMC with estimate b for n steps and step size h. Note that this auxiliary process is defined only for purposes of analysis; it cannot be used for practical algorithm as we do not have access to ∇f.

We can couple this process with LMC using s so that as long as X_t does not hit B, the processes agree, thus satisfying condition 1 of Theorem 3.7.1.

Then by Chebyshev's inequality,

$$P(B) \leqslant \left(\frac{\varepsilon}{\varepsilon_1}\right)^2 =: \delta.$$

Let $T = Nh$. Then by Theorem 2.4.2,

$$\chi^2(\tilde{q}_{kh}||p) \leqslant \exp\left(-\frac{kh}{4C_{\mathrm{LS}}}\right)\chi^2(q_0||p) + 680dL^2hC_{\mathrm{LS}} + 20\varepsilon_1^2 C_{\mathrm{LS}} \leqslant \exp\left(-\frac{kh}{4C_{\mathrm{LS}}}\right)\chi^2(q_0||p) + 1.$$

For this to be bounded by ε_χ^2, it suffices for the terms to be bounded by $\frac{\varepsilon_\chi^2}{2}, \frac{\varepsilon_\chi^2}{4}, \frac{\varepsilon_\chi^2}{4}$; this is implied by

$$T \geqslant 4C_{\mathrm{LS}} \ln\left(\frac{2K_\chi}{\varepsilon_\chi^2}\right) =: T_{\min}$$

$$h = \frac{\varepsilon_\chi^2}{4392dL^2C_{\mathrm{LS}}}$$

$$\varepsilon_1 = \frac{\varepsilon_\chi}{4\sqrt{5C_{\mathrm{LS}}}}.$$

(We choose h so that the condition in Theorem 2.4.2 is satisfied; note $\varepsilon_\chi \leqslant 1$.) By Theo-

rem 2.4.1,

$$\begin{aligned}
\mathrm{TV}(q_{Nh}, \overleftarrow{q}_{Nh}) &\leqslant \sum_{k=0}^{N-1} (1 + \chi^2(q_{kh}||p))^{1/2} P(B)^{1/2} \\
&\leqslant \left(\sum_{k=0}^{N-1} \exp\left(-\frac{kh}{8C_{\mathrm{LS}}}\right) \chi^2(q_0||p)^{1/2} + 2 \right) \delta^{1/2} \\
&\leqslant \left(\left(\sum_{k=0}^{\infty} \exp\left(-\frac{kh}{8C_{\mathrm{LS}}}\right) K_\chi \right) + 2N \right) \frac{\varepsilon}{\varepsilon_1} \\
&\leqslant \frac{\varepsilon}{\varepsilon_1} \left(\frac{16 C_{\mathrm{LS}}}{h} K_\chi + 2N \right).
\end{aligned}$$

In order for this to be $\leqslant \varepsilon_{\mathrm{TV}}$, it suffices for

$$\varepsilon \leqslant \varepsilon_1 \varepsilon_{\mathrm{TV}} \left(\frac{1}{4N} \wedge \frac{h}{32 C_{\mathrm{LS}} K_\chi} \right).$$

Supposing that we run for time T where $T_{\min} \leqslant T \leqslant C_T T_{\min}$, we have that $N = \frac{T}{h} \leqslant \frac{C_T T_{\min}}{h}$. Thus it suffices for

$$\begin{aligned}
\varepsilon &\leqslant \varepsilon_1 \varepsilon_{\mathrm{TV}} \left(\frac{h}{4 C_T T_{\min}} \wedge \frac{h}{32 C_{\mathrm{LS}} K_\chi} \right) \\
&= \frac{\varepsilon_\chi}{4\sqrt{5 C_{\mathrm{LS}}}} \cdot \varepsilon_{\mathrm{TV}} \cdot \frac{\varepsilon_\chi^2}{2720 d L^2 C_{\mathrm{LS}}} \left(\frac{1}{16 C_T C_{\mathrm{LS}} \ln(2K_\chi/\varepsilon_\chi^2)} \wedge \frac{1}{32 C_{\mathrm{LS}} K_\chi} \right) \\
&= \frac{\varepsilon_{\mathrm{TV}} \varepsilon_\chi^3}{174080 \sqrt{5} d L^2 C_{\mathrm{LS}}^{5/2} (C_T \ln(2K_\chi/\varepsilon_\chi^2) \vee 2K_\chi)}.
\end{aligned}$$

□

2.6.3 Proof of Theorem 2.2.2

We restate the theorem for convenience.

Theorem 2.2.2 (Annealed LMC with L^2-accurate score estimate). *Let $p : \mathbb{R}^d \to \mathbb{R}$ be a probability density satisfying Assumption 1 for $M_1 = O(d)$, and let $p_{\sigma^2} := p * \varphi_{\sigma^2}$. Suppose furthermore that $\nabla \ln p_{\sigma^2}$ is L-Lipschitz for every $\sigma \geqslant 0$. Given $\sigma_{\min} > 0$, there exists a sequence*

$\sigma_{\min} = \sigma_1 < \cdots < \sigma_M$ *with* $M = O\left(\sqrt{d}\log\left(\frac{dC_{\mathrm{LS}}}{\sigma_{\min}^2}\right)\right)$ *such that for each* m, *if*

$$\left\|\nabla \ln(p_{\sigma_m^2}) - s(\cdot, \sigma_m^2)\right\|^2_{L^2(p_{\sigma_m^2})} = \mathbb{E}_{p_{\sigma_m^2}}[\left\|\nabla \ln p_{\sigma_m^2}(x) - s(x, \sigma_m^2)\right\|^2] \leqslant \varepsilon^2.$$

$$\text{with } \varepsilon := \widetilde{O}\left(\frac{\varepsilon_{\mathrm{TV}}^{4.5}}{d^{3.25}L^2C_{\mathrm{LS}}^{2.5}}\right) \tag{2.3}$$

then $x^{(1)}$ *is a sample from a distribution* q *such that* $\mathrm{TV}(q, p_{\sigma_1^2}) \leqslant \varepsilon_{\mathrm{TV}}$.

Proof. We choose

$$h_M = \cdots = h_2 = \Theta\left(\frac{1}{dL^2C_{\mathrm{LS}}}\right) \qquad h_1 = \Theta\left(\frac{dL^2C_{\mathrm{LS}}}{\varepsilon_{\mathrm{TV}}^2}\right)$$

$$T_{M-1} = \cdots = T_2 = \Theta\left(C_{\mathrm{LS}}\ln\left(\frac{M}{\varepsilon_{\mathrm{TV}}}\right)\right) \qquad T_1 = \Theta\left(C_{\mathrm{LS}}\ln\left(\frac{1}{\varepsilon_{\mathrm{TV}}}\right)\right),$$

and $T_M = 0$, $N_m = T_m/h$.

Choose the sequence $\sigma_{\min}^2 = \sigma_1^2 < \cdots < \sigma_M^2$ to be geometric with ratio $1 + \Theta\left(\frac{1}{\sqrt{d}}\right)$. Note that

$$\chi^2(N(0, \sigma_2^2 I_d)||N(0, \sigma_1^2 I_d)) = \frac{\sigma_1^d}{\sigma_2^{2d}}(2\sigma_2^{-2} - \sigma_1^{-2})^{-d/2} - 1 = \left(\frac{\sigma_2^2}{\sigma_1^2}\right)^{-d/2}\left(2 - \left(\frac{\sigma_2}{\sigma_1}\right)^2\right)^{-\frac{d}{2}}.$$

For $\sigma_2^2 = (1+\varepsilon)\sigma_1^2$, this equals $(1+\varepsilon)^{-d/2}(1-\varepsilon)^{-d/2} = (1-\varepsilon^2)^{-d/2} - 1$. For $\varepsilon = \Theta\left(\frac{1}{\sqrt{d}}\right)$, this is $d \cdot O\left(\frac{1}{d}\right) = O(1)$. Hence, the χ^2-divergence between successive distributions $p_{\sigma_m^2}$ is $O(1)$. Choosing $\sigma_M^2 = \Omega(d(M_1 + C_{\mathrm{LS}}))$ ensures we have a warm start for the highest noise level by Lemma 2.9.9: $\chi^2(p_{\mathrm{prior}}||p_{\sigma_M^2}) = O(1)$. This uses $O\left(\sqrt{d}\log\left(\frac{dC_{\mathrm{LS}}}{\sigma_{\min}^2}\right)\right)$ noise levels.

Write $p_m = p_{\sigma_m^2}$ for short. Let q_m be the distribution of the final sample $x^{(m)}$. We show by downwards induction on m that there is $\overline{q}_m$ such that

$$\mathrm{TV}(q_m, \overline{q}_m) \leqslant \frac{(M+1) - m}{M+1}\varepsilon_{\mathrm{TV}}$$

$$\chi^2(\overline{q}_m||p_m) \leqslant \left(\frac{\varepsilon_{\mathrm{TV}}}{4(M+1)}\right)^2.$$

For $m = M$, this follows from the assumption on ε and Theorem 2.2.1 with $K_\chi = O(1)$ (given by the warm start).

Fix $m < M$ and suppose it holds for $m+1$. We use the closeness between q_{m+1} and p_{m+1} combined with $\chi^2(p_{m+1}||p_m) = O(1)$ to obtain compute how close q_{m+1} and p_m are. Because the triangle inequality does not hold for χ^2, we will incur an extra TV error.

Let $\overline{q}_{m,m+1}$ be the distribution of the final sample if $x_0^{(m+1)} \sim \overline{q}_m$. We have

$$\mathrm{TV}(q_{m+1}, \overline{q}_{m,m+1}) \leqslant \mathrm{TV}(q_m, \overline{q}_m) \leqslant \frac{(M+1)-m}{M+1}\varepsilon_{\mathrm{TV}}.$$

By Markov's inequality, when $\chi^2(p_{m+1}||p_m) \leqslant 1$,

$$\mathbb{P}_{p_{m+1}}\left(\frac{p_{m+1}}{p_m} \geqslant \frac{8(M+1)}{\varepsilon_{\mathrm{TV}}}\right) \leqslant \frac{\chi^2(p_{m+1}||p_m)+1}{8(M+1)/\varepsilon_{\mathrm{TV}}} \leqslant \frac{\varepsilon_{\mathrm{TV}}}{4(M+1)}.$$

Let $\overline{q}_{m+1,m} = \mathbb{1}_{\left\{\frac{p_{m+1}}{p_m} \leqslant \frac{8(M+1)}{\varepsilon_{\mathrm{TV}}}\right\}}\overline{q}_{m+1} \Big/ \int_{\left\{\frac{p_{m+1}}{p_m} \leqslant \frac{8(M+1)}{\varepsilon_{\mathrm{TV}}}\right\}} \overline{q}_{m+1}$. Note that (using $\mathrm{TV}(\overline{q}_{m+1}, p_{m+1}) \leqslant \sqrt{\chi^2(\overline{q}_{m+1}||p_{m+1})} \leqslant \frac{\varepsilon_{\mathrm{TV}}}{4(M+1)}$)

$$\begin{aligned}\mathbb{P}_{\overline{q}_{m+1}}\left(\frac{p_{m+1}}{p_m} \geqslant \frac{8(M+1)}{\varepsilon_{\mathrm{TV}}}\right) &\leqslant \mathbb{P}_{p_{m+1}}\left(\frac{p_{m+1}}{p_m} \geqslant \frac{8(M+1)}{\varepsilon_{\mathrm{TV}}}\right) + \mathrm{TV}(\overline{q}_{m+1}, p_{m+1}) \\ &\leqslant \frac{\varepsilon_{\mathrm{TV}}}{4(M+1)} + \frac{\varepsilon_{\mathrm{TV}}}{4(M+1)} \leqslant \frac{1}{2}. \end{aligned} \tag{2.19}$$

so $\overline{q}_{m+1,m} \leqslant 2\overline{q}_{m+1}$ and

$$\begin{aligned}\chi^2(\overline{q}_{m+1,m}||p_m) + 1 &\leqslant 2(\chi^2(\overline{q}_{m+1}||p_m)+1) \\ &= \int_{\left\{\frac{p_{m+1}}{p_m} \leqslant \frac{8(M+1)}{\varepsilon_{\mathrm{TV}}}\right\}} \frac{\overline{q}_{m+1}(x)^2}{p_{m+1}(x)^2} \cdot \frac{p_{m+1}(x)}{p_m(x)} p_{m+1}(x)\,dx \\ &\leqslant \frac{8(M+1)}{\varepsilon_{\mathrm{TV}}}(\chi^2(\overline{q}_{m+1}||p_{m+1})+1) \leqslant \frac{16(M+1)}{\varepsilon_{\mathrm{TV}}}.\end{aligned}$$

Let $\overline{q}'_{m+1,m}$ be the distribution of $x_{N_m}^{(m)}$ when $x_0^{(m)} \sim \overline{q}_{m+1,m}$. Then by assumption on ε (2.3) and Theorem 2.2.1 (with $K_\chi = 4\sqrt{\frac{M+1}{\varepsilon_{\mathrm{TV}}}}$, $\varepsilon_\chi = \frac{\varepsilon_{\mathrm{TV}}}{4(M+1)}$, and $\varepsilon_{\mathrm{TV}} \leftarrow \frac{\varepsilon_{\mathrm{TV}}}{2(M+1)}$), there is $\overline{q}_m$ such

that $\mathrm{TV}(\overline{q}'_{m,m+1}, \overline{q}_m) \leqslant \frac{\varepsilon_{\mathrm{TV}}}{2(M+1)}$ and $\chi^2(\overline{q}_m || p_m) \leqslant \frac{\varepsilon_{\mathrm{TV}}}{4(M+1)}$. It remains to bound

$$\begin{aligned}
\mathrm{TV}(q_m, \overline{q}_m) &\leqslant \mathrm{TV}(q_m, \overline{q}'_{m,m+1}) + \mathrm{TV}(\overline{q}'_{m,m+1}, \overline{q}_m) \\
&\leqslant \mathrm{TV}(q_{m+1}, \overline{q}_{m,m+1}) + \frac{\varepsilon_{\mathrm{TV}}}{2(M+1)} \\
&\leqslant \mathrm{TV}(q_{m+1}, \overline{q}_{m+1}) + \mathrm{TV}(\overline{q}_{m+1}, \overline{q}_{m+1,m}) + \frac{\varepsilon_{\mathrm{TV}}}{2(M+1)} \\
&\leqslant \frac{(M+1)-(m+1)}{M+1}\varepsilon_{\mathrm{TV}} + \mathbb{P}_{\overline{q}_{m+1}}\left(\frac{p_{m+1}}{p_m} \geqslant \frac{8(M+1)}{\varepsilon_{\mathrm{TV}}}\right) + \frac{\varepsilon_{\mathrm{TV}}}{2(M+1)} \\
&\leqslant \frac{(M+1)-(m+1)}{M+1}\varepsilon_{\mathrm{TV}} + \frac{\varepsilon_{\mathrm{TV}}}{2(M+1)} + \frac{\varepsilon_{\mathrm{TV}}}{2(M+1)} = \frac{(M+1)-m}{M+1}\varepsilon_{\mathrm{TV}},
\end{aligned}$$

where we use (2.19) in the last line. This finishes the induction step.

Finally, the theorem follows by taking $m = 1$ and noting

$$\begin{aligned}
\mathrm{TV}(q_1, p_1) &\leqslant \mathrm{TV}(q_1, \overline{q}_1) + \mathrm{TV}(\overline{q}_1, p_1) \\
&\leqslant \mathrm{TV}(q_1, \overline{q}_1) + \sqrt{\chi^2(\overline{q}_1 || p_1)} \leqslant \frac{M\varepsilon_{\mathrm{TV}}}{M+1} + \frac{\varepsilon_{\mathrm{TV}}}{4(M+1)} \leqslant \varepsilon_{\mathrm{TV}}.
\end{aligned}$$

□

2.7 Analysis for SGM based on reverse SDE's

In this section, we analyze score-based generative models based on reverse SDE's. In Section 2.7.2, we prove convergence of the predictor algorithm under L^∞-accurate score estimate (Theorem 2.4.3, restated as 2.7.1) using lemmas proved in Section 2.7.3, 2.7.4, 2.7.5, and 2.7.6. In Section 2.7.7, we prove convergence of the predictor algorithm under L^2-accurate score estimate (Theorem 2.3.1, restated as 2.7.16). In Section 2.7.8, we prove convergence of the predictor-corrector algorithm (Theorem 2.3.2).

2.7.1 Discretization and Score Estimation

With a change of variable in (2.4), we define the sampling process x_t on $[0, T]$ by

$$dx_t = [-f(x_t, T-t) + g(T-t)^2 \nabla \ln \tilde{p}_{T-t}(x_t)]\, dt + g(T-t)\, dw_t, \;\; x_0 \sim \tilde{p}_T.$$

Denoting the distribution of x_t by p_t and running the process from 0 to T, we will exactly obtain $p_T = \tilde{p}_0$, which is the data distribution. In practice, we need to discretize this

process and replace the score function $\nabla \ln \tilde{p}_{T-t}$ with the estimated score s. With a general Euler-Maruyama method, we would obtain $\{z_k\}_{k=0}^N$ defined by

$$z_{(k+1)h} = z_{kh} - h \cdot [f(z_{kh}, T-kh) - g(T-kh)^2 s(z_{kh}, T-kh)] + \sqrt{h} \cdot g(T-kh)\eta_{k+1}, \quad (2.20)$$

where $h = T/N$ is the step size and η_k is a sequence of independent Gaussian random vectors. As we run (2.20) from 0 to N with h small enough, we should expect that the distribution of z_T is close to that of x_T. However, in both SMLD or DDPM models, for fixed z_k, the integration

$$\int_{kh}^{(k+1)h} f(z_{kh}, T-t)\, dt \text{ and } s(z_{kh}, T-kh) \cdot \int_{kh}^{(k+1)h} g(T-t)^2\, dt$$

can be exactly computed, as can the diffusion term. Therefore, we can consider the following process z_t as an "interpolation" of (2.20):

$$dz_t = [-f(z_{kh}, T-t) + g(T-t)^2 s(z_{kh}, T-kh)]\, dt + g(T-t)\, dw_t, \quad t \in [kh, (k+1)h]. \quad (2.21)$$

Note that by running this process instead, we can reduce the discretization error. Now if we denote the distribution of z_t by q_t, with $q_0 \approx p_0$, we can expect that q_T is close to p_T. Here the estimated score s satisfies for all x

$$\|s(x, T-kh) - \nabla \ln \tilde{p}_{T-kh}(x)\| \leqslant \varepsilon_{kh}, \quad k = 0, 1, \ldots, N. \quad (2.22)$$

Observe that in either SMLD or DDPM, the function $g(t)^2$ is Lipschitz on $[0, T]$. So in the following sections, we will assume that $g(t)^2$ is L_g-Lipschitz on $[0, T]$.

2.7.2 Predictor

In this section, we present the main result (Theorem 2.7.1) on the one-step error of the predictor in χ^2-divergence, which can be obtained by directly applying the Gronwall's inequality to the differential inequality derived in Lemma 2.7.3. Note that Theorem 2.7.1 is a more precise version of Theorem 2.4.3; see the remark following the theorem.

Theorem 2.7.1. *With the setting in Section 2.7.1, assume g is non-decreasing and let*

$$h \leqslant \min_{kh \leqslant t \leqslant (k+1)h} \frac{1}{g(T-kh)^2(28L^2 + 10C_t + \mathbb{E}_{p_t}\|x\|^2 + 64C_{t,L} + 128C_{d,L} + 360L_s^2(\tilde{R}_t + 2C_t R_d))}$$

be positive, where C_t is the log-Sobolev constant of p_t, bounded in Lemma 2.9.7. Suppose that $\nabla \ln p_t$ is L-Lipschitz for all $t \in [kh, (k+1)h]$, $s(\cdot, kh)$ is L_s-Lipschitz, $L, L_s \geqslant 1$, and ε_{kh} is such that (2.22) *holds. Then*

$$\chi^2(q_{(k+1)h} || p_{(k+1)h}) \leqslant \left[\chi^2(q_{kh} || p_{kh}) + \int_{kh}^{(k+1)h} C_{t,kh}\, dt\right] e^{\int_{kh}^{(k+1)h} (-\frac{1}{8C_t} + 8\varepsilon_{kh}^2) g(T-t)^2\, dt}$$

Here,

$$C_{t,kh} = \left[8\varepsilon_{kh}^2 + E \cdot (t - kh) g(T - kh)^2\right] g(T-t)^2$$

and

$$E = 9(4L_s^2 + 1) + 8C_{d,L}$$

$$C_{t,L} = \begin{cases} 32L^2 & \text{in SMLD,} \\ (88C_t^2 + 400)L^2 & \text{in DDPM,} \end{cases}$$

$$C_{d,L} = \begin{cases} 76L^2 d & \text{in SMLD,} \\ 6 + 94L^2 d & \text{in DDPM} \end{cases} \leqslant 100L^2 d$$

$$\tilde{R}_t = 9(C_t + 1)$$

$$R_d = 300d + 12$$

are defined in (2.24), (2.27), (2.28), (2.30) *and* (2.31), *respectively.*

Proof. The theorem follows from applying Gronwall's inequality to the result of Lemma 2.7.3. □

Remark. Note that in DDPM, $E = O(L_s^2 + L^2 d)$. Therefore, when $g \equiv 1$, $C_{t,kh} = O(\varepsilon_1^2 + (L_s^2 + L^2 d)h)$, where we denote the upper bound of ε_{kh} for all $k \in \{0, ..., N\}$ by ε_1. Using the bound on the log-Sobolev constant (Lemma 2.9.7) and second moment (Lemma 2.9.8) for DDPM, we note that the restriction on h for all steps is implied by

$$h = O\left(\frac{1}{\mathbb{E}_{p_{\text{data}}} \|x\|^2 + C_{\text{LS}}(C_{\text{LS}} + d)(L \vee L_s)^2}\right)$$

with appropriate constants. Then we can conclude the first inequality in Theorem 2.4.3 by combining Theorem 2.7.1 and Lemma 2.9.7 and the second inequality from unfolding the

first one and evaluating the geometric series. Likewise, we have the following analogue for SMLD, for which we omit the proof.

Theorem 2.7.2 (Predictor steps under L^∞ bound on score estimate, SMLD)**.** *Let $p : \mathbb{R}^d \to \mathbb{R}$ be a probability density satisfying Assumption 1 and $s(\cdot, t) : \mathbb{R}^d \to \mathbb{R}^d$ be a score estimate s with error bounded in L^∞ for each $t \in [0, T]$:*

$$\|\nabla \ln \widetilde{p}_t - s(\cdot, t)\|_\infty = \max_{x \in \mathbb{R}^d} \|\nabla \ln \widetilde{p}_t(x) - s(x, t)\|] \leqslant \varepsilon_1.$$

Consider SMLD. Let $C_T = C_{\mathrm{LS}} + T$. Let $g \equiv 1$, $T \geqslant C_{\mathrm{LS}} d$, and $h = O\left(\frac{1}{\mathbb{E}_{p_0}\|x\|^2 + C_T d(L \vee L_s)^2}\right)$. Then

$$\chi^2(q_{(k+1)h} || p_{(k+1)h}) \leqslant \chi^2(q_{kh} || p_{kh}) e^{(-\frac{1}{8C_{T-kh}} + 8\varepsilon_1^2)h} + O(\varepsilon_1^2 h + (L_s^2 + L^2 d)h^2)$$

and letting $t = T - Nh$, if $\varepsilon_1 < \frac{1}{128 C_T}$,

$$\chi^2(q_{Nh} || p_{Nh}) \leqslant \left(\frac{C_{\mathrm{LS}} + t}{C_{\mathrm{LS}} + T}\right)^{\frac{1}{16}} \chi^2(q_0 || p_0) + O\left(\ln\left(\frac{C_{\mathrm{LS}} + T}{C_{\mathrm{LS}} + t}\right)(\varepsilon_1^2 + (L_s^2 + L^2 d)h)\right).$$

Moreover, for $q_0 = p_{\mathrm{prior}}$, $q_0 = \varphi_T$, $\chi^2(q_0 || p_0) \leqslant \frac{C_{\mathrm{LS}} d}{T}$.

Remark. We note that in a sense SMLD and DDPM are equivalent, as we can get from one to the other by rescaling in time and space. First we recall that, as discussed in Section 2.3, all the SMLD models are equivalent under rescaling in time. Therefore we can assume $g(t) = e^{t/2}$ and consider the forward SDE for SMLD

$$dx_t = e^{t/2} dw_t,$$

where w_t is a standard Brownian Motion. Now let $y_t = e^{-t/2} x_t$; then

$$dy_t = -\frac{1}{2} y_t dt + dw_t,$$

which is exactly DDPM with $g(t) = 1$. Note that Theorem 2.7.2 uses a different parameterization for SMLD and the resulting complexity is slightly worse.

2.7.3 Differential Inequality

Now we prove a differential inequality involving $\chi^2(q_t||p_t)$. As in (Chewi et al., 2021), the key difficulty is to bound the discretization error. We decompose it into two error terms and bound them in Lemma 2.7.4 and Lemma 2.7.5 separately.

In the following, we will let

$$G_{kh,t} := \int_{kh}^{t} g(T-s)^2 \, ds. \tag{2.23}$$

Lemma 2.7.3. *Let $(q_t)_{0\leqslant t\leqslant T}$ denote the law of the interpolation* (2.21)*. With the setting in Lemma 2.7.1, we have for $t \in [kh, (k+1)h]$,*

$$\frac{d}{dt}\chi^2(q_t||p_t) \leqslant g(T-t)^2 \left[\left(-\frac{1}{8C_t} + 8\varepsilon_{kh}^2\right)\chi^2(q_t||p_t) + \left[8\varepsilon_{kh}^2 + E\cdot(t-kh)g(T-kh)^2\right]\right],$$

where C_t is the LSI constant of p_t, ε_{kh} is the L^∞-score estimation error at time kh and E is defined in (2.24)*.*

Proof. By Lemma 2.5.2 with

$$\widehat{f}(z_{kh}, t) \leftarrow -f(z_{kh}, T-t) + g(T-t)^2 s(z_{kh}, T-kh)$$
$$f(z,t) \leftarrow -f(z, T-t) + g(T-t)^2 \nabla \ln \widetilde{p}_{T-t}(z),$$

we have

$$\begin{aligned}
\frac{d}{dt}\chi^2(q_t||p_t) &= -g(T-t)^2 \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 2\mathbb{E}\Big[\Big\langle \left(-f(z_{kh}, T-t) + g(T-t)^2 s(z_{kh}, T-kh)\right) \\
&\qquad - \left(-f(z, T-t) + g(T-t)^2 \nabla \ln \widetilde{p}_{T-t}(z)\right), \nabla \frac{q_t}{p_t}\Big\rangle\Big] \\
&= -g(T-t)^2 \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 2\mathbb{E}\left[\left\langle f(z_t, T-t) - f(z_{kh}, T-t), \nabla \frac{q_t(z_t)}{p_t(z_t)}\right\rangle\right] \\
&\quad + 2g(T-t)^2 \mathbb{E}\left[\left\langle s(z_{kh}, T-kh) - \nabla \ln \tilde{p}_{T-t}(z_t), \nabla \frac{q_t(z_t)}{p_t(z_t)}\right\rangle\right] \\
&=: -g(T-t)^2 \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + A + B.
\end{aligned}$$

By Lemma 2.7.4,

$$A \leqslant g(T-t)^2 \left[2(\chi^2(q_t||p_t)+1)\mathbb{E}\left[\|z_t - z_{kh}\|^2 \psi_t(z_t) \right] + \frac{1}{8}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) \right],$$

while by Lemma 2.7.5,

$$\begin{aligned} B &\leqslant \frac{1}{2} g(T-t)^2 \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 8g(T-t)^2 L_s^2 (\chi^2(q_t||p_t)+1)\mathbb{E}\left[\|z_t - z_{kh}\|^2 \psi_t(z_t) \right] \\ &\quad + 8\left[\varepsilon_{kh}^2 + G_{kh,t} C_{d,L}\right] g(T-t)^2 (\chi^2(q_t||p_t)+1). \end{aligned}$$

Therefore, for $h \leqslant \frac{1}{72g(T-kh)^2(4L_s^2+1)(\tilde{R}_t \vee 2C_t R_d)} \wedge \frac{1}{128 g(T-kh)^2 C_{d,L}}$, using Lemma 2.7.15,

$$\begin{aligned} \frac{d}{dt}\chi^2(q_t||p_t) &\leqslant -\frac{3}{8} g(T-t)^2 \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 8\left[\varepsilon_{kh}^2 + G_{kh,t} C_{d,L}\right] g(T-t)^2 (\chi^2(q_t||p_t)+1) \\ &\quad + g(T-t)^2 (8L_s^2+2)(\chi^2(q_t||p_t)+1)\mathbb{E}\left[\|z_t - z_{kh}\|^2 \psi_t(z_t) \right] \\ &\leqslant -\frac{3}{8} g(T-t)^2 \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) \\ &\quad + 9g(T-t)^2 (4L_s^2+1) G_{kh,t}\left[\tilde{R}_t \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + R_{t,kh}(\chi^2(q_t||p_t)+1) \right] \\ &\quad + 8\left[\varepsilon_{kh}^2 + G_{kh,t} C_{d,L}\right] g(T-t)^2 (\chi^2(q_t||p_t)+1) \\ &\leqslant -\frac{2}{8} g(T-t)^2 \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + g(T-t)^2 \frac{1}{8C_t}\chi^2(q_t||p_t) + 8g(T-t)^2 \varepsilon_{kh}^2 \chi^2(q_t||p_t) \\ &\quad + g(T-t)^2 \left[8\varepsilon_{kh}^2 + 8C_{d,L} G_{kh,t} + 9(4L_s^2+1) G_{kh,t} R_d \right]. \end{aligned}$$

Using the fact that p_t satisfies a log-Sobolev inequality with constant C_t,

$$\begin{aligned} \frac{d}{dt}\chi^2(q_t||p_t) &\leqslant -\frac{2}{8C_t} g(T-t)^2 \chi^2(q_t||p_t) + \frac{1}{8C_t} g(T-t)^2 \chi^2(q_t||p_t) + 8g(T-t)^2 \varepsilon_{kh}^2 \chi^2(q_t||p_t) \\ &\quad + g(T-t)^2 \left[8\varepsilon_{kh}^2 + 8C_{d,L} G_{kh,t} + 9(4L_s^2+1) G_{kh,t} R_d \right] \\ &\leqslant \left(-\frac{1}{8C_t} + 8\varepsilon_{kh}^2 \right) g(T-t)^2 \chi^2(q_t||p_t) + g(T-t)^2 [8\varepsilon_{kh}^2 + E(t-kh) g(T-kh)^2]. \end{aligned}$$

where

$$E = 9(4L_s^2+1) + 8C_{d,L}. \tag{2.24}$$

□

In order to bound the error terms A and B, we will use Lemma 2.5.3. Let $\phi_t(x) = \frac{q_t(x)}{p_t(x)}$ and $\psi_t(x) = \phi_t(x)/\mathbb{E}_{p_t}\phi_t^2$. Then $\mathbb{E}\psi_t(z_t) = 1$ and in fact the normalizing factor $\mathbb{E}_{p_t}\phi_t^2 = \chi^2(q_t||p_t) + 1$. We first deal with error term A.

Lemma 2.7.4. *In the setting of Lemma 2.7.3, we have the following bound for term A:*

$$2\mathbb{E}\left[\left\langle f(z_t, T-t) - f(z_{kh}, T-t), \nabla \frac{q_t(z_t)}{p_t(z_t)} \right\rangle\right]$$
$$\leqslant g(T-t)^2 \left[2(\chi^2(q_t||p_t) + 1)\mathbb{E}\left[\|z_t - z_{kh}\|^2 \psi_t(z_t)\right] + \frac{1}{8}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right)\right].$$

Proof. In SMLD, $f(x,t) = 0$ and hence $A = 0$; while in DDPM, $f(x,t) = -\frac{1}{2}g(t)^2 x$. Therefore, by Lemma 2.5.3,

$$2\mathbb{E}\left[\left\langle f(z_t, T-t) - f(z_{kh}, T-t), \nabla \frac{q_t(z_t)}{p_t(z_t)} \right\rangle\right]$$
$$= -g(T-t)^2\mathbb{E}\left[\left\langle z_t - z_{kh}, \nabla \frac{q_t(z_t)}{p_t(z_t)} \right\rangle\right]$$
$$\leqslant g(T-t)^2 \left[2 \cdot \mathbb{E}_{p_t}\phi_t^2 \cdot \mathbb{E}\left[\|z_t - z_{kh}\|^2 \psi_t(z_t)\right] + \frac{1}{8}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right)\right]$$
$$= g(T-t)^2 \left[2(\chi^2(q_t||p_t) + 1)\mathbb{E}\left[\|z_t - z_{kh}\|^2 \psi_t(z_t)\right] + \frac{1}{8}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right)\right]. \qquad \square$$

Now we bound error term B.

Lemma 2.7.5. *In the setting of Lemma 2.7.3, we have the following bound for term B:*

$$B \leqslant \frac{1}{2}g(T-t)^2\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 8g(T-t)^2 L_s^2(\chi^2(q_t||p_t) + 1)\mathbb{E}\left[\|z_t - z_{kh}\|^2 \psi_t(z_t)\right]$$
$$+ 8\left[\varepsilon_{kh}^2 + G_{kh,t}C_{d,L}\right] g(T-t)^2(\chi^2(q_t||p_t) + 1).$$

Proof. We first decompose the error:

$$\begin{aligned}
&\mathbb{E}\left[\left\langle s(z_{kh}, T-kh) - \nabla \ln \tilde{p}_{T-t}(z_t), \nabla \frac{q_t(z_t)}{p_t(z_t)} \right\rangle\right] \\
&= \mathbb{E}\left[\left\langle s(z_{kh}, T-kh) - s(z_t, T-kh), \nabla \frac{q_t(z_t)}{p_t(z_t)} \right\rangle\right] \\
&\quad + \mathbb{E}\left[\left\langle s(z_t, T-kh) - \nabla \ln p_{kh}(z_t), \nabla \frac{q_t(z_t)}{p_t(z_t)} \right\rangle\right] \\
&\quad + \mathbb{E}\left[\left\langle \nabla \ln p_{kh}(z_t) - \nabla \ln p_t(z_t), \nabla \frac{q_t(z_t)}{p_t(z_t)} \right\rangle\right] \\
&=: B_1 + B_2 + B_3.
\end{aligned}$$

Now we bound these error terms separately. For B_1, by the Lipschitz assumption, we have by Lemma 2.5.3, for a constant $C_2 > 0$ to be chosen later,

$$\begin{aligned}
B_1 &\leqslant \mathbb{E}\left[L_s \left\|z_{kh} - z_t\right\| \cdot \left\|\nabla \frac{q_t(z_t)}{p_t(z_t)}\right\|\right] \\
&\leqslant 4L_s^2 \cdot \mathbb{E}_{p_t}\phi_t^2 \cdot \mathbb{E}\left[\left\|z_t - z_{kh}\right\|^2 \psi_t(z_t)\right] + \frac{1}{16}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) \\
&= 4L_s^2(\chi^2(q_t||p_t) + 1)\mathbb{E}\left[\left\|z_t - z_{kh}\right\|^2 \psi_t(z_t)\right] + \frac{1}{16}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right).
\end{aligned}$$

For B_2, recalling the assumption that $\|s(x, T-kh) - \nabla \ln p_{kh}(x)\| \leqslant \varepsilon_{kh}$ for all x, we have by Lemma 2.5.3

$$\begin{aligned}
B_2 &\leqslant 4\mathbb{E}\left[\left\|s(z_t, T-kh) - \nabla \ln p_{kh}(z_t)\right\|^2 \psi_t(z_t)\right] \cdot \mathbb{E}_{p_t}[\phi_t^2] + \frac{1}{16}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) \\
&\leqslant 4\varepsilon_{kh}^2(\chi^2(q_t||p_t) + 1) + \frac{1}{16}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right).
\end{aligned} \tag{2.25}$$

Now for the last error term B_3, we have by Lemma 2.5.3 that

$$\begin{aligned}
B_3 &\leqslant 4\mathbb{E}_{p_t}\phi_t^2 \cdot \mathbb{E}\left[\left\|\nabla \ln p_{kh}(z_t) - \nabla \ln p_t(z_t)\right\|^2 \psi_t(z_t)\right] + \frac{1}{16}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) \\
&\leqslant 4K_{t,kh}(\chi^2(q_t||p_t) + 1) + \frac{1}{16}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right).
\end{aligned} \tag{2.26}$$

Here $K_{t,kh}$ is the bound for $\mathbb{E}\left[\psi_t(z_t)\left\|\nabla \ln p_{kh}(z_t) - \nabla \ln p_t(z_t)\right\|^2\right]$ obtained in Lemma 2.7.13:

$$K_{t,kh} := G_{kh,t}\left[\frac{C_{t,L}}{\chi^2(q_t||p_t)+1}\cdot \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + C_{d,L}\right]$$

where $C_{t,L}$ and $C_{d,L}$ are constants defined in (2.27) and (2.28) respectively. Hence

$$B_3 \leqslant 4G_{kh,t}\left[C_{t,L}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + C_{d,L}(\chi^2(q_t||p_t)+1)\right] + \frac{1}{16}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right).$$

Combining all these results, we finally obtain the bound for error term B in Lemma 2.7.3: for $h \leqslant \frac{1}{64C_{t,L}g(T-kh)^2}$,

$$\begin{aligned}
B &= 2g(T-t)^2(B_1+B_2+B_3)\\
&\leqslant \frac{3}{8}g(T-t)^2\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 8g(T-t)^2L_s^2(\chi^2(q_t||p_t)+1)\mathbb{E}\left[\|z_t - z_{kh}\|^2\psi_t(z_t)\right]\\
&\quad + 8C_{t,L}g(T-t)^2G_{kh,t}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right)\\
&\quad + 8\left[\varepsilon_{kh}^2 + G_{kh,t}C_{d,L}\right]g(T-t)^2(\chi^2(q_t||p_t)+1)\\
&\leqslant \frac{1}{2}g(T-t)^2\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 8g(T-t)^2L_s^2(\chi^2(q_t||p_t)+1)\mathbb{E}\left[\|z_t - z_{kh}\|^2\psi_t(z_t)\right]\\
&\quad + 8\left[\varepsilon_{kh}^2 + G_{kh,t}C_{d,L}\right]g(T-t)^2(\chi^2(q_t||p_t)+1). \qquad \square
\end{aligned}$$

2.7.4 Change of Measure

As shown in Lemma 2.7.4 and Lemma 2.7.5, the key to the proof of Lemma 2.7.3 is bounding the discretization error A and B. The difficulty is that these errors usually have the form of $\mathbb{E}_{\psi_t q_t}\left[\|u(x)\|^2\right]$ for some function $u : \mathbb{R}^d \to \mathbb{R}^d$, while it is usually easier to bound those expectations over the original probability measure or our target distribution p_t. Therefore, as discussed in (Chewi et al., 2021, Section 5.1), our task is to bound these error terms under a complicated change of measure. We first state such a result with respect to the gradient of the potential.

Lemma 2.7.6. *(Chewi et al., 2021, Lemma 16) Assume that $p(x) \propto e^{-V(x)}$ is a density in $\mathbb{R}^d$ and*

$\nabla V(x)$ is L-Lipschitz. Then for any probability density q, it holds that

$$\mathbb{E}_q\left[\|\nabla V\|^2\right] \leqslant 4\mathbb{E}_p\left[\left\|\nabla\sqrt{\frac{q(x)}{p(x)}}\right\|^2\right] + 2dL = \mathbb{E}_q\left[\left\|\nabla \ln\frac{q(x)}{p(x)}\right\|^2\right] + 2dL.$$

Proof. Define the Langevin diffusion w.r.t. $p(x)$:

$$dx_t = -\nabla V(x_t)\,dt + \sqrt{2}\,dw_t,$$

where B_t is a standard Brownian Motion in $\mathbb{R}^d$. Let $\mathcal{L}$ be the corresponding infinitesimal generator, i.e., $\mathcal{L}f = \langle\nabla V, \nabla f\rangle - \Delta f$. Observe that $\mathcal{L}V = \|\nabla V\|^2 - \Delta V$ and $\mathbb{E}_p\mathcal{L}f = 0$ for any f, so

$$\begin{aligned}
\mathbb{E}_q\left[\|\nabla V\|^2\right] &= \mathbb{E}_q\mathcal{L}V + \mathbb{E}_q\Delta V \\
&\leqslant \int \mathcal{L}V\left(\frac{q(x)}{p(x)} - 1\right)p(x)dx + dL = \int\left\langle\nabla V, \nabla\frac{q(x)}{p(x)}\right\rangle p(x)dx + dL \\
&= 2\int\left\langle\sqrt{\frac{q(x)}{p(x)}}\nabla V, \nabla\sqrt{\frac{q(x)}{p(x)}}\right\rangle p(x)dx + dL \\
&\leqslant \frac{1}{2}\mathbb{E}_q\left[\|\nabla V\|^2\right] + 2\mathbb{E}_p\left[\left\|\nabla\sqrt{\frac{q(x)}{p(x)}}\right\|^2\right] + dL.
\end{aligned}$$

Rearrange this inequality to obtain the desired result. □

Now applying this Lemma to $p = p_t$ and $q = \psi_t q_t$, we get immediately the following corollary. Note that $\psi_t q_t$ is a density function because $\int \psi_t(x)q_t(x)\,dx = 1$ and $\psi_t(x)q_t(x) \geqslant 0$ for any $x \in \mathbb{R}^d$.

Corollary 2.7.7. *In the setting of Lemma 2.7.3, it holds that*

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

Proof. Applying Lemma 2.7.6 to the density $\psi_t q_t$ yields

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

□

Note that we cannot expect analogous results for a general $u(x)$ as in Lemma 2.7.6. In the general case, we apply the Donsker-Varadhan variational principle, which states that for probability measures p and q,

$$\mathbb{E}_q \|u(x)\|^2 \leqslant \mathrm{KL}(q||p) + \ln \mathbb{E}_p \exp \|u(x)\|^2.$$

Towards this end, we first need to analyze $\mathrm{KL}(\psi_t q_t||p_t)$.

Lemma 2.7.8. *Let $\phi_t(x) = \frac{q_t(x)}{p_t(x)}$ and $\psi_t(x) = \phi_t(x)/\mathbb{E}_{p_t}\phi_t^2$. If p_t satisfies a LSI with constant C_t, then*

$$\mathrm{KL}(\psi_t q_t||p_t) \leqslant \frac{2C_t}{\chi^2(q_t||p_t)+1} \cdot \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right).$$

Proof. Since p_t satisfies LSI with constant C_t,

$$\begin{aligned}
\mathrm{KL}(\psi_t q_t||p_t) &\leqslant \frac{C_t}{2}\int \left\|\nabla \ln \frac{\psi_t(x)q_t(x)}{p_t(x)}\right\|^2 \psi_t(x)q_t(x)dx \\
&= 2C_t \int \left\|\nabla \ln \frac{q_t(x)}{p_t(x)}\right\|^2 \psi_t(x)q_t(x)dx \\
&= 2C_t \int \left\|\nabla \frac{q_t(x)}{p_t(x)}\right\|^2 \frac{\psi_t(x)p_t(x)^2}{q_t(x)}dx \\
&= \frac{2C_t}{\chi^2(q_t||p_t)+1} \cdot \int \left\|\nabla \frac{q_t(x)}{p_t(x)}\right\|^2 p_t(x)dx \\
&= \frac{2C_t}{\chi^2(q_t||p_t)+1} \cdot \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right).
\end{aligned}$$

□

With this in hand, we are ready to bound the second moment of $\psi_t q_t$ as well as the variance of a Gaussian random vector with respect to this measure:

Lemma 2.7.9. *With the setting of Lemma 2.7.3, we have*

$$\mathbb{E}\left[\psi_t(z_t)\|z_t\|^2\right] \leqslant \frac{2C_t^2}{\chi^2(q_t||p_t)+1} \cdot \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + \frac{1}{2}\mathbb{E}_{p_t}\left[\|x\|^2\right] + \frac{1}{2}C_t,$$

where C_t is the LSI constant of p_t, which is bounded in Lemma 2.9.6, and the second moment of p_t is bounded in Lemma 2.9.8.

Proof. Since p_t has LSI constant C_t, by Donsker-Varadhan variational principle,

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

for any $s > 0$. By Lemma 2.9.1, for any $s \in [0, \frac{1}{C_t})$, we have

$$\mathbb{E}_{p_t}\left[e^{\frac{s}{2}\left\|x\right\|^2}\right] \leqslant \frac{1}{\sqrt{1 - C_t \cdot s}} \exp\left[\frac{s}{2(1 - C_t \cdot s)}(\mathbb{E}_{p_t}\left\|x\right\|)^2\right].$$

Now choose $s = \frac{1}{2C_t}$, we have

$$\mathbb{E}_{p_t}\left[e^{\frac{s}{2}\left\|x\right\|^2}\right] \leqslant \sqrt{2}\exp\left[\frac{1}{2C_t}(\mathbb{E}_{p_t}\left\|x\right\|)^2\right].$$

Hence

$$\mathbb{E}\left[\psi_t(z_t)\left\|z_t\right\|^2\right] \leqslant C_t \cdot \left[\mathrm{KL}(\psi_t q_t||p_t) + \frac{1}{2C_t}\mathbb{E}_{p_t}\left[\left\|x\right\|^2\right] + \frac{\ln 2}{2}\right].$$

Now with the bound of $\mathrm{KL}(\psi_t q_t||p_t)$ in Lemma 3.5.1, we obtain

$$\mathbb{E}\left[\psi_t(z_t)\left\|z_t\right\|^2\right] \leqslant \frac{2C_t^2}{\chi^2(q_t||p_t) + 1} \cdot \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + \frac{1}{2}\mathbb{E}_{p_t}\left[\left\|x\right\|^2\right] + \frac{1}{2}C_t. \qquad \square$$

Lemma 2.7.10. *With the setting of Lemma 2.7.3,*

$$\mathbb{E}\left[\psi_t(z_t)\left\|\int_{kh}^t g(T-s)dw_s\right\|^2\right] \leqslant 2\int_{kh}^t g(T-s)^2\,ds \cdot \left[\frac{8C_t}{\chi^2(q_t||p_t) + 1} \cdot \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + d + 8\ln 2\right],$$

where C_t is the LSI constant of p_t.

Proof. Note that $\int_{kh}^t g(T-s)dw_s$ is a Gaussian random vector with variance $\int_{kh}^t g(T-s)^2 ds \cdot I_d$. Using the Donsker-Varadhan variational principle, for any random variable X,

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

Applying this to $X = c\left(\left\|\int_{kh}^t g(T-s)dw_s\right\| - \mathbb{E}\left\|\int_{kh}^t g(T-s)dw_s\right\|\right)^2$ for a constant $c > 0$ to be chosen later, we can bound

$$\tilde{\mathbb{E}}\left\|\int_{kh}^t g(T-s)dw_s\right\|^2 \leqslant 2\mathbb{E}\left[\left\|\int_{kh}^t g(T-s)dw_s\right\|^2\right] + \frac{2}{c}\left[\mathrm{KL}(\tilde{\mathbb{P}}||\mathbb{P}) + Q_{t,kh}\right],$$

where $\frac{d\tilde{\mathbb{P}}}{d\mathbb{P}} = \psi_t(z_t)$ and

$$Q_{t,kh} = \ln \mathbb{E} \exp\left(c\Big(\left\|\int_{kh}^{t} g(T-s)dw_s\right\| - \mathbb{E}\left\|\int_{kh}^{t} g(T-s)dw_s\right\|\Big)^2\right).$$

Now following (Chewi et al., 2021, Theorem 4), we set $c = \frac{1}{8\int_{kh}^{t} g(s)^2 ds}$, so that

$$\mathbb{E} \exp\left[\frac{\left(\left\|\int_{kh}^{t} g(T-s)dw_s\right\| - \mathbb{E}\left\|\int_{kh}^{t} g(T-s)dw_s\right\|\right)^2}{8\int_{kh}^{t} g(s)^2 ds}\right] \leqslant 2.$$

Next, using the LSI for p_t, 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 $\mathbb{E}_{p_t}\phi_t^2 = \chi^2(q_t||p_t) + 1 \geqslant 1$, we have that

$$\mathrm{KL}(\tilde{\mathbb{P}}||\mathbb{P}) \leqslant \frac{1}{2}\mathrm{KL}(\psi_t q_t||p_t) \leqslant \frac{C_t}{\chi^2(q_t||p_t)+1} \cdot \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right),$$

where the last inequality is due to Lemma 3.5.1. We have proved

$$\begin{aligned}&\mathbb{E}\left[\psi_t(z_t)\left\|\int_{kh}^{t} g(T-s)\, dw_s\right\|^2\right] \\ &\leqslant 2d\int_{kh}^{t} g(T-s)^2\, ds + 16\int_{kh}^{t} g(T-s)^2 ds \cdot \left[\frac{C_t}{\chi^2(q_t||p_t)+1} \cdot \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + \ln 2\right] \\ &\leqslant 2\int_{kh}^{t} g(T-s)^2\, ds \cdot \left[\frac{8C_t}{\chi^2(q_t||p_t)+1} \cdot \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + d + 8\ln 2\right].\end{aligned}$$

□

2.7.5 Perturbation Error

In the previous section, we bound errors in the form of $\mathbb{E}_{\psi_t q_t}\|u(x)\|^2$ with a change of measure technique, where $\|u(x)\|^2$ is easy to bound with respect to the original measure or p_t. However, this is not always the case for the errors we are considering. In this section, we aim to bound $\mathbb{E}_{\psi_t q_t}\left[\|\nabla \ln p_{kh}(x) - \nabla \ln p_t(x)\|^2\right]$, where, as discussed in Lemma 2.7.13,

p_{kh} can be regarded as a perturbed version of p_t with some Gaussian noise. We first provide a point-wise bound for SMLD (Lemma 2.7.11) and DDMP (Lemma 2.7.12), respectively and then use them to bound the expectation with respect to $\psi_t q_t$.

Lemma 2.7.11. *Suppose that $p(x) \propto e^{-V(x)}$ is a probability density on $\mathbb{R}^d$, where $V(x)$ is L-smooth, and let $\varphi_{\sigma^2}(x)$ be the density function of $N(0, \sigma^2 I_d)$. Then for $L \leqslant \frac{1}{2\sigma^2}$,*

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

Proof. Note that

$$\nabla \ln p * \varphi_{\sigma^2}(x) = \frac{\int_{\mathbb{R}^d} -\nabla V(y) e^{-V(y)} e^{-\frac{\|x-y\|^2}{2\sigma^2}} \, dy}{\int_{\mathbb{R}^d} e^{-V(y)} e^{-\frac{\|x-y\|^2}{2\sigma^2}} \, dy} = -\mathbb{E}_{p_{x,\sigma^2}} \nabla V(y),$$

where p_{x,σ^2} denotes the probability density

$$p_{x,\sigma^2}(y) \propto p(y) e^{-\frac{\|y-x\|^2}{2\sigma^2}}$$

so when V is L-smooth,

$$\begin{aligned} \left\| \nabla \ln \frac{p(x)}{p * \varphi_{\sigma^2}(x)} \right\| &= \left\| \mathbb{E}_{p_{x,\sigma^2}} [\nabla V(y) - \nabla V(x)] \right\| \\ &\leqslant \mathbb{E}_{p_{x,\sigma^2}} [L \left\| y - x \right\|] \end{aligned}$$

We now write

$$\mathbb{E}_{p_{x,\sigma^2}} \left\| y - x \right\| \leqslant \mathbb{E}_{p_{x,\sigma^2}} \left\| y - \mathbb{E}_{p_{x,\sigma^2}} y \right\| + \left\| \mathbb{E}_{p_{x,\sigma^2}} y - y^* \right\| + \left\| y^* - x \right\|,$$

where $y^* \in \operatorname{argmax}_y p_{x,\sigma^2}(y)$ is a mode of the distribution p_{x,σ^2}. We now bound each of these terms.

1. For the first term, note that p_{x,σ^2} is $\left(\frac{1}{\sigma^2} - L\right)$-strongly convex, so satisfies a Poincaré inequality with constant $\left(\frac{1}{\sigma^2} - L\right)^{-1}$. Thus

$$\begin{aligned} \mathbb{E}_{p_{x,\sigma^2}} \left\| y - x \right\| &\leqslant \mathbb{E}_{p_{x,\sigma^2}} [\left\| y - \mathbb{E}_{p_{x,\sigma^2}} y \right\|^2]^{1/2} \\ &= \left(\sum_{i=1}^{d} \operatorname{Var}_{p_{x,\sigma^2}}(y_i) \right)^{1/2} \leqslant \left(d \left(\frac{1}{\sigma^2} - L \right)^{-1} \right)^{1/2}. \end{aligned}$$

2. For the second term, by Lemma 2.9.3, noting that $V(y) + \frac{\|x-y\|^2}{2\sigma^2}$ is $\left(\frac{1}{\sigma^2} + L\right)$-smooth,

$$\begin{aligned}
\left\|\mathbb{E}_{p_{x,\sigma^2}} y - y^*\right\| &\leqslant \left(\frac{1}{\sigma^2} - L\right)^{-1/2} d^{1/2} \left(5 + \ln\left(\left(\frac{1}{\sigma^2} - L\right)^{-1}\left(\frac{1}{\sigma^2} + L\right)\right)\right)^{1/2} \\
&\leqslant \left(\frac{1}{\sigma^2} - L\right)^{-1/2} d^{1/2} \left(5 + \ln \frac{1 + L\sigma^2}{1 - L\sigma^2}\right)^{1/2} \\
&\leqslant \sqrt{7}\left(\frac{1}{\sigma^2} - L\right)^{-1/2} d^{1/2},
\end{aligned}$$

   where the last inequality uses $\sigma^2 \leqslant \frac{1}{2L}$.

3. For the third term, we note that the mode satisfies

$$\begin{aligned}
\nabla V(y^*) + \frac{y^* - x}{\sigma^2} &= 0 \\
-\frac{y^* - x}{\sigma^2} &= \nabla V(y^*) = (\nabla V(y^*) - \nabla V(x)) + \nabla V(x) \\
\frac{1}{\sigma^2}\|y^* - x\| &\leqslant \|\nabla V(x)\| + L\|y^* - x\| \\
\|y^* - x\| &\leqslant \left(\frac{1}{\sigma^2} - L\right)^{-1} \|\nabla V(x)\|.
\end{aligned}$$

Putting these together and using $\left(\frac{1}{\sigma^2} - L\right)^{-1} \leqslant 2$, we obtain

$$\begin{aligned}
\left\|\nabla \ln \frac{p(x)}{p * \varphi_{\sigma^2}(x)}\right\| &\leqslant (\sqrt{7} + 1)L\left(\frac{1}{\sigma^2} - L\right)^{-1/2} d^{1/2} + L\left(\frac{1}{\sigma^2} - L\right)^{-1} \|\nabla V(x)\| \\
&\leqslant 6L\sigma d^{1/2} + 2L\sigma^2 \|\nabla V(x)\|.
\end{aligned}$$

□

Lemma 2.7.12. *With the setting in Lemma 2.7.11 and the notation $p_\alpha(x) = \alpha^d p(\alpha x)$ for $\alpha \geqslant 1$, we have that for $L \leqslant \frac{1}{2\alpha^2\sigma^2}$,*

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

Proof. Note $p_\alpha(x)$ is also a probability density in $\mathbb{R}^d$. By the triangle inequality,

$$\left\|\nabla \ln \frac{p(x)}{(p_\alpha * \varphi_{\sigma^2})(x)}\right\| \leqslant \left\|\nabla \ln \frac{p(x)}{p_\alpha(x)}\right\| + \left\|\nabla \ln \frac{p_\alpha(x)}{(p_\alpha * \varphi_{\sigma^2})(x)}\right\|.$$

Without loss of generality, we can assume that $p(x) = e^{-V(x)}$; then $p_\alpha(x) = \alpha^d e^{-V(\alpha x)}$. Hence

$$\begin{aligned}\left\|\nabla \ln \frac{p(x)}{p_\alpha(x)}\right\| &= \|\alpha \nabla V(\alpha x) - \nabla V(x)\| \\ &\leqslant \|\alpha \nabla V(\alpha x) - \alpha \nabla V(x)\| + \|\alpha \nabla V(x) - \nabla V(x)\| \\ &\leqslant \alpha(\alpha-1)L\|x\| + (\alpha-1)\|\nabla V(x)\|.\end{aligned}$$

Since $\alpha \nabla V(\alpha x)$ is $\alpha^2 L$-Lipschitz, by Lemma 2.7.11,

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

By the Lipschitz assumption,

$$\|\nabla V(\alpha x)\| \leqslant \|\nabla V(\alpha x) - \nabla V(x)\| + \|\nabla V(x)\| \leqslant (\alpha-1)L\|x\| + \|\nabla V(x)\|.$$

The result follows from combining the three inequalities above. □

Lemma 2.7.13. *In the setting of Lemma 2.7.3, we have for $t \in [kh, (k+1)h]$,*

$$\begin{aligned}&\mathbb{E}\left[\psi_t(z_t)\|\nabla \ln p_{kh}(z_t) - \nabla \ln p_t(z_t)\|^2\right] \\ &\leqslant G_{kh,t} \cdot \left[\frac{C_{t,L}}{\chi^2(q_t||p_t)+1} G_{kh,t} \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + C_{d,L}\right],\end{aligned}$$

where

$$C_{t,L} = \begin{cases} 32L^2 & \text{in SMLD,} \\ (88C_t^2 + 400)L^2 & \text{in DDPM,} \end{cases} \tag{2.27}$$

and

$$C_{d,L} = \begin{cases} 76L^2 d & \text{in SMLD,} \\ 6 + 94L^2 d & \text{in DDPM} \end{cases} \leqslant 100L^2 d. \tag{2.28}$$

Proof. In both SMLD and DDPM models, we have the following relationship for $t \in [kh, (k+1)h]$:

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

where $p_\alpha(x) = \alpha^d p(\alpha x)$. In SMLD, $\alpha = 1$ and $\sigma^2 = \int_{kh}^t g(T-s)^2\,ds$, while in DDPM, $\alpha = e^{\frac{1}{2}\int_{kh}^t g(T-s)^2\,ds}$ and $\sigma^2 = 1 - e^{-\int_{kh}^t g(T-s)^2\,ds}$. Now for SMLD,

$$
\begin{aligned}
&\mathbb{E}\left[\psi_t(z_t)\left\|\nabla \ln p_{kh}(z_t) - \nabla \ln p_t(z_t)\right\|^2\right] \\
&\leqslant 72L^2\sigma^2 d + 8L^2\sigma^4 \mathbb{E}\left[\psi_t(z_t)\left\|\nabla \ln p_t(z_t)\right\|^2\right] && \text{by Lemma 2.7.11} \\
&\leqslant 72\sigma^2 L^2 d + \frac{32L^2\sigma^4}{\chi^2(q_t||p_t)+1}\cdot \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 16\sigma^4 L^3 d && \text{by Corollary 2.7.7} \\
&\leqslant G_{kh,t}^2\left(\frac{32L^2}{\chi^2(q_t||p_t)+1} + 16L^3 d\right) + G_{kh,t}\cdot 72L^2 d \\
&\leqslant G_{kh,t}^2 \frac{32L^2}{\chi^2(q_t||p_t)+1} + G_{kh,t}\cdot 76L^2 d,
\end{aligned}
$$

where in the last inequality we use the fact that g is increasing, so that for $h \leqslant \frac{1}{4Lg(T-kh)^2}$,

$$
G_{kh,t}L = \int_{kh}^t g(T-s)^2\,ds \cdot L \leqslant h\cdot g(T-kh)^2\cdot L \leqslant \frac{1}{4}.
$$

Recall that to use Lemma 2.7.11, it suffices that $L \leqslant \frac{1}{2\alpha^2\sigma^2}$, and so it suffices that $h \leqslant \frac{1}{4Lg(T-kh)^2}$ in SMLD.

For DDPM, observe that for $h \leqslant \frac{1}{4g(T-kh)^2}$,

$$
\begin{aligned}
\alpha &\leqslant 1 + \int_{kh}^t g(T-s)^2 ds \leqslant 1 + (t-kh)g(T-kh)^2 \leqslant 1 + \frac{1}{4} \\
\sigma^2 &= 1 - e^{-\int_{kh}^t g(T-s)^2 ds} \leqslant \int_{kh}^t g(T-s)^2 ds \leqslant (t-kh)g(T-kh)^2 \leqslant \frac{1}{4}.
\end{aligned}
$$

By Lemma 3.4.7, using the assumption that $L \geqslant 1$, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\psi_t(z_t)\left\|\nabla \ln p_{kh}(z_t) - \nabla \ln p_t(z_t)\right\|^2\right] \\
&\leqslant 72\alpha^4 L^2\sigma^2 d + 4(\alpha + 2\alpha^3 L\sigma^2)^2(\alpha-1)^2 L^2 \mathbb{E}\left[\psi(z_t)\left\|z_t\right\|^2\right] \\
&\quad + 4(\alpha - 1 + 2\alpha^3 L\sigma^2)^2 \mathbb{E}\left[\psi_t(z_t)\left\|\nabla \ln p_t(z_t)\right\|^2\right] \\
&\leqslant 72\alpha^4 L^2\sigma^2 d + 44L^2 G_{kh,t}^2 \mathbb{E}\left[\psi(z_t)\left\|z_t\right\|^2\right] \\
&\quad + 100L^2 G_{kh,t}^2 \mathbb{E}\left[\psi_t(z_t)\left\|\nabla \ln p_t(z_t)\right\|^2\right] \\
&\leqslant 44L^2 d G_{kh,t} \\
&\quad + 44L^2\left[\frac{2C_t^2}{\chi^2(q_t||p_t)+1}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + \frac{1}{2}\mathbb{E}_{p_t}\left\|x\right\|^2 + \frac{1}{2}C_t\right] G_{kh,t}^2 \\
&\quad + 100L^2\left[\frac{4}{\chi^2(q_t||p_t)+1}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 2dL\right] G_{kh,t}^2 \\
&\leqslant L^2 G_{kh,t}\Bigg[G_{kh,t}\left(\frac{88C_t^2+400}{\chi^2(q_t||p_t)+1}\right)\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) \\
&\quad + 44d + G_{kh,t}\left(22(\mathbb{E}_{p_t}\left\|x\right\|^2 + C_t) + 200Ld\right)\Bigg] \\
&\leqslant G_{kh,t}\left[G_{kh,t}\frac{88C_t^2+400}{\chi^2(q_t||p_t)+1}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 6 + 94L^2 d\right],
\end{aligned}
$$

where we used Lemma 2.7.9 and Corollary 2.7.7. Here, we use the assumption that $h \leqslant \frac{1}{4g(T-kh)^2(\mathbb{E}_{p_t}\left\|x\right\|^2+C_t)}$. □

2.7.6 Auxiliary Lemmas

In this section, we continue with bounding errors in the form of $\mathbb{E}_{\psi_t q_t}\left\|u(x)\right\|^2$. However, we only decompose them into errors which we have already bounded in the previous two sections. The following two lemmas will be directly applied in the proof of Lemma 2.7.4 and Lemma 2.7.5.

Lemma 2.7.14. *With the setting of Lemma 2.7.3, we have the following bound of the second moment*

of estimated score function with respect to $\psi_t q_t$:

$$\mathbb{E}\left[\psi_t(z_t)\left\|s(z_t, T-kh)\right\|^2\right] \leqslant \frac{4C_{t,L}G_{kh,t}+8}{\chi^2(q_t||p_t)+1}\cdot \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 4(\varepsilon_{kh}^2 + C_{d,L} + dL),$$

where $C_{t,L}$ and $C_{d,L}$ are constants defined in Lemma 2.7.13.

Proof. Note that by the triangle inequality,

$$\begin{aligned}\|s(x, T-kh)\| \leqslant \|s(x, T-kh) - \nabla \ln \tilde{p}_{T-kh}(x)\| \\ + \|\nabla \ln \tilde{p}_{T-kh}(x) - \nabla \ln \tilde{p}_{T-t}(x)\| + \|\nabla \ln \tilde{p}_{T-t}(x)\|,\end{aligned}$$

and hence,

$$\begin{aligned}\|s(x, T-kh)\|^2 \leqslant 4\|s(x, T-kh) - \nabla \ln \tilde{p}_{T-kh}(x)\|^2 \\ + 4\|\nabla \ln \tilde{p}_{T-kh}(x) - \nabla \ln \tilde{p}_{T-t}(x)\|^2 + 2\|\nabla \ln \tilde{p}_{T-t}(x)\|^2.\end{aligned}$$

Recall that we need to bound this second moment of estimated score function with respect to $\psi_t q_T$. For the first term, as $\|s(x.T-kh) - \nabla \ln p_{kh}(x)\|$ is ε_{kh}-bounded, we have trivial bound that

$$\mathbb{E}_{\psi_t q_t}\|s(x, T-kh) - \nabla \ln \tilde{p}_{T-kh}(x)\|^2 \leqslant \varepsilon_{kh}^2.$$

By Lemma 2.7.13, the second term is bounded by

$$\begin{aligned}&\mathbb{E}_{\psi_t q_t}\left[\|\nabla \ln p_{kh}(z_t) - \nabla \ln p_t(z_t)\|^2\right] \\ &\leqslant G_{kh,t}\cdot\left[\frac{C_{t,L}}{\chi^2(q_t||p_t)+1}G_{kh,t}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + C_{d,L}\right]\end{aligned}$$

for constant $C_{t,L}$ and $C_{d,L}$ defined in (2.27) and (2.28) respectively. The last term is bounded in Corollary 2.7.7 by

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

Combining these three inequalities, we obtain that for $h \leqslant \frac{1}{g(T-kh)^2}$,

$$\begin{aligned}&\mathbb{E}\left[\psi_t(z_t)\left\|s(z_t, T-kh)\right\|^2\right] \\ &\leqslant \frac{4C_{t,L}+8}{\chi^2(q_t||p_t)+1}G_{kh,t}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 4(\varepsilon_{kh}^2 + C_{d,L} + dL).\end{aligned}$$

□

Now we bound $\mathbb{E}\left[\psi_t(z_t)\left\|z_t - z_{kh}\right\|^2\right]$.

Lemma 2.7.15. *In the setting of Lemma 2.7.3, if*

$$h \leqslant \frac{1}{g(T-kh)^2(8L^2+20L+3L_s+10C_t+\mathbb{E}_{p_t}\|x\|^2)},$$

then

$$\mathbb{E}\left[\psi_t(z_t)\left\|z_t - z_{kh}\right\|^2\right] \leqslant \frac{9}{2}G_{kh,t}\left[\frac{\tilde{R}_t}{\chi^2(q_t||p_t)+1}\cdot \mathscr{E}_{p_t}^{\circ}\left(\frac{q_t}{p_t}\right) + R_{t,kh}\right],$$

where $\tilde{R}_t$ and R_d are defined in (2.30) *and* (2.31) *respectively.*

Proof. Note that

$$\begin{aligned}
&\|z_t - z_{kh}\| \\
&= \left\|G_{kh,t}s(z_{kh}, T-kh) - \int_{kh}^{t} f(z_{kh}, T-s)ds + \int_{kh}^{t} g(T-s)dw_s\right\| \\
&\leqslant G_{kh,t}\|s(z_{kh}, T-kh)\| + \frac{1}{2}\left\|z_{kh}\int_{kh}^{t} g(T-s)^2 ds\right\| + \left\|\int_{kh}^{t} g(T-s)dw_s\right\| \\
&\leqslant G_{kh,t}\left[\|s(z_{kh}, T-kh)\| + \frac{1}{2}\|z_{kh}\|\right] + \left\|\int_{kh}^{t} g(T-s)dw_s\right\| \\
&\leqslant G_{kh,t}\left[\|s(z_t, T-kh)\| + L_s\|z_t - z_{kh}\| + \frac{1}{2}\|z_t\| + \frac{1}{2}\|z_t - z_{kh}\|\right] + \left\|\int_{kh}^{t} g(T-s)dw_s\right\| \\
&= G_{kh,t}\left[\|s(z_t, T-kh)\| + \frac{1}{2}\|z_t\|\right] + \left(L_s + \frac{1}{2}\right)g(T-kh)^2\cdot h\|z_t - z_{kh}\| + \left\|\int_{kh}^{t} g(T-s)dw_s\right\|,
\end{aligned}$$

where the next-to-last line is due to the fact that the estimated score function is L_s-Lipschitz. We also use the fact that $g(t)$ is an increasing function and hence $g(T-t) \leqslant g(T-kh)$ for any $t \in [kh, (k+1)h]$. Hence if $h \leqslant \frac{1}{3(L_s+1/2)g(T-kh)^2}$, then

$$\|z_t - z_{kh}\| \leqslant \frac{3}{2}G_{kh,t}\left[\|s(z_t, T-kh)\| + \frac{1}{2}\|z_t\|\right] + \frac{3}{2}\left\|\int_{kh}^{t} g(T-s)dw_s\right\|.$$

Therefore, by the fact that $(a+b)^2 \leqslant 2a^2 + 2b^2$ for any $a, b > 0$,

$$\|z_t - z_{kh}\|^2 \leqslant \frac{9}{2}G_{kh,t}^2\left[2\|s(z_t, T-kh)\|^2 + \frac{1}{2}\|z_t\|^2\right] + \frac{9}{2}\left\|\int_{kh}^{t} g(T-s)dw_s\right\|^2. \tag{2.29}$$

With the results of Lemma 2.7.14 and Lemma 2.7.9, we have

$$2\mathbb{E}\left[\psi_t(z_t)\left\|s(z_t, T-kh)\right\|^2\right] + \frac{1}{2}\mathbb{E}\left[\psi_t(z_t)\left\|z_t\right\|^2\right]$$

$$\leqslant \frac{8C_{t,L}G_{kh,t} + C_t^2 + 16}{\chi^2(q_t||p_t)+1} \cdot \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 8(\varepsilon_{kh}^2 + C_{d,L} + dL) + \frac{1}{4}\mathbb{E}_{p_t}\left\|x\right\|^2 + \frac{1}{4}C_t.$$

Now plugging this and the result of Lemma 2.7.10 into (2.29), we get that

$$\begin{aligned}\mathbb{E}\left[\psi_t(z_t)\left\|z_t - z_{kh}\right\|^2\right] &\leqslant \frac{9}{2}G_{kh,t}^2 \cdot 8(\varepsilon_{kh}^2 + C_{d,L} + dL) \\ &+ \frac{9}{2}G_{kh,t}^2 \cdot \left[\frac{8C_{t,L}G_{kh,t} + C_t^2 + 16}{\chi^2(q_t||p_t)+1} \cdot \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + \frac{1}{4}\mathbb{E}_{p_t}\left\|x\right\|^2 + \frac{1}{4}C_t\right] \\ &+ 9G_{kh,t} \cdot \left[\frac{8C_t}{\chi^2(q_t||p_t)+1} \cdot \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + d + 8\ln 2\right].\end{aligned}$$

Hence, using the assumption on h,

$$\mathbb{E}\left[\psi_t(z_t)\left\|z_t - z_{kh}\right\|^2\right] \leqslant \frac{9}{2}G_{kh,t}\left[\frac{K_1}{\chi^2(q_t||p_t)+1} \cdot \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + K_2\right],$$

where

$$\begin{aligned}K_1 &:= 8C_{t,L}G_{kh,t}^2 + (C_t^2+16)G_{kh,t} + 16C_t \\ &\leqslant 8(88C_t^2 + 400L^2)\frac{1}{400L + 100C_t^2} + (C_t^2+16)\frac{1}{20L+10C_t} + 8C_t \\ &\leqslant 8 + C_t + 1 + 8C_t = 9(C_t+1)\end{aligned}$$

and

$$\begin{aligned}K_2 &:= \left[\frac{1}{4}(\mathbb{E}_{p_t}\left\|x\right\|^2 + C_t) + 8(\varepsilon_{kh}^2 + C_{d,L} + dL)\right]G_{kh,t} + 2d + 16\ln 2 \\ &\leqslant \left[\frac{1}{4}(\mathbb{E}_{p_t}\left\|x\right\|^2 + C_t) + 8(\varepsilon_{kh}^2 + 256L^2d + dL)\right]\left(\frac{1}{\mathbb{E}_{p_t}\left\|x\right\|^2 + C_t + 8L^2}\right) + 2d + 16\ln 2 \\ &\leqslant \frac{1}{4} + 300d + 16\ln 2 \leqslant 300d + 12.\end{aligned}$$

Hence the lemma holds by setting

$$\tilde{R}_t = 9(C_t+1), \tag{2.30}$$

$$R_d = 300d + 12. \tag{2.31}$$

□

2.7.7 Proof of Theorem 2.3.1

We state a more precise version of Theorem 2.3.1. The structure of the proof is similar to that of Theorem 2.2.1.

Theorem 2.7.16 (Predictor with L^2-accurate score estimate, DDPM)**.** *Let $p_{\text{data}} : \mathbb{R}^d \to \mathbb{R}$ be a probability density satisfying Assumption 1 with $M_2 = O(d)$, and let $\widetilde{p}_t$ be the distribution resulting from evolving the forward SDE according to DDPM with $g \equiv 1$. Suppose furthermore that $\nabla \ln \widetilde{p}_t$ is L-Lipschitz for every $t \geqslant 0$, and that each $s(\cdot, t)$ satisfies Assumption 3. Then if*

$$\varepsilon = O\left(\frac{\varepsilon_{\text{TV}}\varepsilon_\chi^3}{(C_{\text{LS}} + d)C_{\text{LS}}^{5/2}(L \vee L_s)^2(\ln(C_{\text{LS}}d) \vee C_{\text{LS}}\ln(1/\varepsilon_{\text{TV}}^2))}\right),$$

running (P) *starting from p_{prior} for time $T = \Theta\left(\ln(C_{\text{LS}}d) \vee C_{\text{LS}}\ln\left(\frac{1}{\varepsilon_{\text{TV}}}\right)\right)$ and step size $h = \Theta\left(\frac{\varepsilon_\chi^2}{C_{\text{LS}}(C_{\text{LS}}+d)(L\vee L_s)^2}\right)$ results in a distribution q_T such that q_T is ε_{TV}-far in TV distance from a distribution $\overline{q}_T$, where $\overline{q}_T$ satisfies $\chi^2(\overline{q}_T||p_{\text{data}}) \leqslant \varepsilon_\chi^2$. In particular, taking $\varepsilon_\chi = \varepsilon_{\text{TV}}$, we have $\text{TV}(q_T||p_{\text{data}}) \leqslant 2\varepsilon_{\text{TV}}$.*

Proof of Theorem 2.7.16. We first define the bad sets where the error in the score estimate is large,

$$B_t := \{\|\nabla \ln p_t(x) - s(x, T-t)\| > \varepsilon_1\} \tag{2.32}$$

for some ε_1 to be chosen.

Given $t \geqslant 0$, let $t_- = h\left\lfloor \frac{t}{h} \right\rfloor$. Given a bad set B, define the interpolated process by

$$d\overline{z}_t = -\left[f(z_{t_-}, T-t) - g(T-t)^2 b(z_{kh}, T-kh)\right]\, dt + g(T-t)\, dw_t, \tag{2.33}$$

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

In other words, simulate the reverse SDE using the score estimate as long as the point is in the good set (for the current p_t) at the previous discretization step, 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 q_{nh}

is the distribution resulting from running LMC with estimate b for n steps and step size h. Note that this process is defined only for purposes of analysis, as we do not have access to $\nabla \ln p_t$.

We can couple this process with the predictor algorithm using s so that as long as $x_{mh} \notin B_{mh}$, the processes agree, thus satisfying condition 1 of Theorem 3.7.1.

Then by Chebyshev's inequality,

$$P(B_t) \leqslant \left(\frac{\varepsilon}{\varepsilon_1}\right)^2 =: \delta.$$

Let $T = Nh$, and let $K_\chi = \chi^2(q_0||p_0)$. Then by Theorem 2.4.3,

$$\begin{aligned}\chi^2(\bar{q}_{kh}||p_{kh}) &= \exp\left(-\frac{kh}{16C_{\mathrm{LS}}}\right)\chi^2(q_0||p_0) + O\left(C_{\mathrm{LS}}(\varepsilon_1^2 + (L_s^2 + L^2 d)h)\right) \\ &= \exp\left(-\frac{kh}{4C_{\mathrm{LS}}}\right)\chi^2(\mu_0||p) + O(1).\end{aligned}$$

For this to be bounded by ε_χ^2, it suffices for the terms to be bounded by $\frac{\varepsilon_\chi^2}{2}, \frac{\varepsilon_\chi^2}{4}, \frac{\varepsilon_\chi^2}{4}$; this is implied by

$$\begin{aligned}T &\geqslant 32C_{\mathrm{LS}} \ln\left(\frac{2K_\chi}{\varepsilon_\chi^2}\right) =: T_{\min} \\ h &= O\left(\frac{\varepsilon_\chi^2}{C_{\mathrm{LS}}(C_{\mathrm{LS}} + d)(L \vee L_s)^2}\right) \\ \varepsilon_1 &= O\left(\frac{\varepsilon_\chi}{\sqrt{C_{\mathrm{LS}}}}\right).\end{aligned}$$

(We choose h so that the condition in Theorem 2.4.3 is satisfied; note $\varepsilon_\chi \leqslant 1$.) By Theo-

rem 3.7.1,

$$\begin{aligned}
\mathrm{TV}(q_{nh}, \bar{q}_{nh}) &\leqslant \sum_{k=0}^{n-1} (1 + \chi^2(q_{kh}||p))^{1/2} P(B_{kh})^{1/2} \\
&\leqslant \left(\sum_{k=0}^{n-1} \exp\left(-\frac{kh}{32 C_{\mathrm{LS}}}\right) \chi^2(q_0||p)^{1/2} + O(1) \right) \delta^{1/2} \\
&\leqslant \left(\left(\sum_{k=0}^{\infty} \exp\left(-\frac{kh}{32 C_{\mathrm{LS}}}\right) K_\chi \right) + O(n) \right) \frac{\varepsilon}{\varepsilon_1} \\
&\leqslant \frac{\varepsilon}{\varepsilon_1} \left(\frac{64 C_{\mathrm{LS}}}{h} K_\chi + O(n) \right).
\end{aligned}$$

In order for this to be $\leqslant \varepsilon_{\mathrm{TV}}$, it suffices for

$$\varepsilon \leqslant \varepsilon_1 \varepsilon_{\mathrm{TV}} \cdot O\left(\frac{1}{n} \wedge \frac{h}{C_{\mathrm{LS}} K_\chi} \right).$$

Supposing that we run for time $T = \Theta(T_{\mathrm{min}})$, we have that $n = \frac{T}{h} = O\left(\frac{C_T T_{\mathrm{min}}}{h}\right)$. Thus it suffices for

$$\begin{aligned}
\varepsilon &= \varepsilon_1 \varepsilon_{\mathrm{TV}} \cdot O\left(\frac{h}{T_{\mathrm{min}}} \wedge \frac{h}{32 C_{\mathrm{LS}} K_\chi} \right) \\
&= O\left(\frac{\varepsilon_\chi}{\sqrt{C_{\mathrm{LS}}}} \cdot \varepsilon_{\mathrm{TV}} \cdot \frac{\varepsilon_\chi^2}{C_{\mathrm{LS}}(C_{\mathrm{LS}} + d)(L \vee L_s)^2} \left(\frac{1}{C_{\mathrm{LS}} \ln(2K_\chi/\varepsilon_\chi^2)} \wedge \frac{1}{C_{\mathrm{LS}} K_\chi} \right) \right) \\
&= O\left(\frac{\varepsilon_{\mathrm{TV}} \varepsilon_\chi^3}{C_{\mathrm{LS}}^{5/2}(C_{\mathrm{LS}} + d)(L \vee L_s)^2 (\ln(2K_\chi/\varepsilon_\chi^2) \vee K_\chi)} \right).
\end{aligned}$$

Finally, note that for $T = \Omega(\ln(C_{\mathrm{LS}} d))$, we have $K_\chi = O(1)$ by Lemma 2.9.9. Substituting $K_\chi = O(1)$ then gives the desired bound. □

2.7.8 Proof of Theorem 2.3.2

We now prove the main theorem on the predictor-corrector algorithm with L^2-accurate score estimate.

Theorem 2.3.2 (Predictor-corrector with L^2-accurate score estimate)**.** *Keep the setup of Theo-*

rem 2.3.1. Then for $\varepsilon_{\mathrm{TV}}^3 = O\left(\frac{1}{(1+L_s/L)^2(1+C_{\mathrm{LS}}/d)(\ln(C_{\mathrm{LS}}d)\vee C_{\mathrm{LS}})}\right)$*, if*

$$\varepsilon = O\left(\frac{\varepsilon_{\mathrm{TV}}^4}{dL^2 C_{\mathrm{LS}}^{5/2}\ln(1/\varepsilon_\chi^2)}\right), \tag{2.5}$$

then Algorithm 2 with appropriate choices of $T = \Theta\left(\ln(C_{\mathrm{LS}}d)\vee C_{\mathrm{LS}}\log\left(\frac{1}{\varepsilon_{\mathrm{TV}}}\right)\right)$*,* N_m*, corrector step sizes* h_m *and predictor step size* h*, produces a sample from a distribution* q_T *such that* $\mathrm{TV}(q_T, p_{\mathrm{data}}) < \varepsilon_{\mathrm{TV}}$.

For simplicity, we consider the predictor-corrector algorithm in the case where all the corrector steps are at the end (but see the discussion following the proof for the general case). The result will follow from chaining together the guarantee on the predictor algorithm (Theorem 2.7.16) and LMC (Theorem 2.2.1).

Proof of Theorem 2.3.2. Let $M = T/h$. We take $h = \Theta\left(\frac{1}{(L\vee L_s)^2 C_{\mathrm{LS}}(C_{\mathrm{LS}}+d)}\right)$, number of corrector steps $N_0 = \cdots = N_{T/h-1} = 0$ and $N_M = T_{\mathrm{c}}/h_M$, where $T_{\mathrm{c}} = \Theta\left(C_{\mathrm{LS}}\ln\left(\frac{2}{\varepsilon_\chi^2}\right)\right)$ and $h_M = \Theta\left(\frac{\varepsilon_\chi^2}{dL^2 C_{\mathrm{LS}}}\right)$. Let the distribution of $z_{T,0}$ be $q_{T,0}$. By Theorem 2.7.16, if $T = \Theta(\ln(C_{\mathrm{LS}}d)\vee C_{\mathrm{LS}}\ln(1/\varepsilon_{\mathrm{TV}}))$, then

$$\varepsilon = O\left(\frac{\varepsilon_{\mathrm{TV}}}{(L\vee L_s)^2(C_{\mathrm{LS}}+d)C_{\mathrm{LS}}^{5/2}(\ln(C_{\mathrm{LS}}d)\vee C_{\mathrm{LS}}\ln(1/\varepsilon_{\mathrm{TV}}))}\right),$$

then there exists $\overline{q}_{T,0}$ such that $\mathrm{TV}(q_{T,0}, \overline{q}_{T,0}) = \varepsilon_{\mathrm{TV}}/2$ and $\chi^2(\overline{q}_{T,0}||p_{\mathrm{data}}) = 1$. Then using Theorem 2.2.1 with $\varepsilon_{\mathrm{TV}} \leftarrow \varepsilon_{\mathrm{TV}}/2$ and $K_\chi = 1$, plus the triangle inequality gives that if

$$\varepsilon = O\left(\frac{\varepsilon_{\mathrm{TV}}\varepsilon_\chi^3}{dL^2 C_{\mathrm{LS}}^{5/2}\ln(1/\varepsilon_{\mathrm{TV}})}\right),$$

then there is $\overline{q}_T$ such that $\mathrm{TV}(q_T, \overline{q}_T) = \varepsilon_{\mathrm{TV}}$ and $\chi^2(\overline{q}_T||p_{\mathrm{data}}) = \varepsilon_\chi^2$. Finally, setting $\varepsilon_{\mathrm{TV}}, \varepsilon_\chi \leftarrow \varepsilon_{\mathrm{TV}}/2$ gives $\mathrm{TV}(q_T, p_{\mathrm{data}}) \leqslant \varepsilon_{\mathrm{TV}}$.

We note that for $\varepsilon_{\mathrm{TV}}^3 = O\left(\frac{1}{(1+L_s/L)^2(1+C_{\mathrm{LS}}/d)(\ln(C_{\mathrm{LS}}d)\vee C_{\mathrm{LS}})}\right)$, the second condition on ε is more constraining, giving the theorem. □

Remark. *We can also analyze a setting where predictor and corrector steps are interleaved; for instance, if $N = 1$, then interleaving the one-step inequalities in Theorem 2.4.2 and 2.4.3 gives a recurrence*

$$\chi^2(q_{(k+1)h,0}||p_{(k+1)h}) \leqslant \exp\left(-\frac{h_{\text{pred}}}{16C_{\text{LS}}}\right)\chi^2(q_{kh,1}||p_{kh}) + O(dL^2h^2 + \varepsilon_1^2 h)$$

$$\chi^2(q_{(k+1)h,1}||p_{(k+1)h)}) \leqslant \exp\left(-\frac{h_{\text{corr}}}{4C_{\text{LS}}}\right)\chi^2(q_{(k+1)h,0}||p_{(k+1)h}) + O(\varepsilon_1^2 h + (L_s^2 + L^2 d)h^2);$$

we can then follow the proof of Theorem 2.3.1. While this does not improve the parameter dependence under the assumptions of Theorem 2.3.2, it can potentially allow for larger step sizes (beyond what is allowed by Theorem 2.3.1), as error accrued in the predictor step can be exponentially damped by the corrector step.

2.8 Stationary distribution of LD with score estimate can be arbitrarily far away

We show that the stationary distribution of Langevin dynamics with L^2-accurate score estimate can be arbitrarily far from the true distribution. We can construct a counterexample even in one dimension, and take the true distribution as a standard Gaussian $p(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$. We will take the score estimate to also be in the form $\nabla \ln q$, so that the stationary distribution of LMC with the score estimate is q. The main idea of the construction is to set q to disagree with p only in the tail of p, where it has a large mode; this error will fail to be detected under $L^2(p)$.

Theorem 2.8.1. *Let p be the density function of $N(0,1)$. There exists an absolute constant C such that given any $\varepsilon > 0$, there exists a distribution q such that*

1. *$\ln q$ is C-smooth.*
2. *$\mathbb{E}_p[\|\nabla \ln p - \nabla \ln q\|^2] < \varepsilon$*
3. *$\text{TV}(p,q) > 1 - \varepsilon$.*

Proof. Take a smooth non-negative function g supported on $[-1,1]$, with $\max|g''| \leqslant c$ and

$g(0) = 1$. We consider a family of distributions for $L > 0$ with density

$$q_L(x) \propto e^{-V_L(x)}, \qquad \text{and} \quad V_L(x) := \frac{x^2}{2} - L^2 g\left(\frac{2}{L}(x - L)\right).$$

Thus the score function for q_L is given by

$$V_L'(x) = x - (2L) g'\Big(\frac{2}{L}(x - L)\Big).$$

We compute the $L^2(p)$ error between the score functions associated with p and q_L.

$$\begin{aligned}
\mathbb{E}_p(V_L'(x) - x)^2 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (2L)^2 \Big| g'\Big(\frac{2}{L}(x - L)\Big)\Big|^2 e^{-x^2/2}\, dx \\
&\leqslant \frac{1}{\sqrt{2\pi}} (2L)^2 e^{-L^2/8} \int_{-\infty}^{\infty} \Big| g'\Big(\frac{2}{L}(x - L)\Big)\Big|^2 dx \\
&= \frac{1}{\sqrt{2\pi}} 2L^3 e^{-L^2/8} \int_{-\infty}^{\infty} \big|g'(y)\big|^2 dy,
\end{aligned}$$

where in the first inequality we have used that $g(\frac{2}{L}(x - L))$ has support $[\frac{L}{2}, \frac{3L}{2}]$, since g has support $[-1, 1]$. Thus the $L^2(p)$-error of the score function goes to 0 as $L \to \infty$.

Moreover, as

$$\big|V_L''(x)\big| = \Big|1 - 4g''\Big(\frac{2}{L}(x - L)\Big)\Big| \leqslant 1 + 4 \max_y \big|g''(y)\big| \leqslant 1 + 4c,$$

the distribution q_L satisfies the required smoothness (Lipschitz score) assumption. Note that q_L has a large mode concentrated at $x = L$ as

$$V_L(L) = \frac{L^2}{2} - L^2 g(0) = -\frac{L^2}{2},$$

while p has vanishing density there, which is in fact the reason that $L^2(p)$-loss of the score estimate is not able to detect the difference between the two distributions. As the height (and width) of the mode becomes arbitrarily large compared to $x = 0$, we have $q_L([\frac{L}{2}, \frac{3L}{2}]) \to 1$, whereas $p_L([\frac{L}{2}, \frac{3L}{2}]) \to 0$. Hence $\mathrm{TV}(p_L, q_L) \to 1$. □

2.9 Useful facts

In this section, we collect some facts and technical lemmas used throughout the paper.

2.9.1 Facts about probability distributions

Given a probability measure P on $\mathbb{R}^d$ with density p, we say that a Poincaré inequality (PI) holds with constant C_P if for any probability measure q,

$$\chi^2(q||p) \leqslant C_P \mathscr{E}_p\left(\frac{q}{p}\right) := C_P \int_{\mathbb{R}^d} \left\|\nabla \frac{q(x)}{p(x)}\right\|^2 p(x)dx. \tag{PI}$$

Alternatively, for any C^1 function f,

$$\mathrm{Var}_p(f) \leqslant C_P \int_{\mathbb{R}^d} \|\nabla f\|^2 \, p(x)\, dx.$$

We say that a log-Sobolev inequality (LSI) holds with constant C_{LS} if for any probability measure q,

$$\mathrm{KL}(q||p) \leqslant \frac{C_{LS}}{2} \int_{\mathbb{R}^d} \left\|\nabla \ln \frac{q(x)}{p(x)}\right\|^2 q(x)dx. \tag{LSI}$$

We call the Poincaré constant and log-Sobolev constant the smallest C_P, C_{LS} for which the inequalities hold for all q. If p satisfies a log-Sobolev inequality with constant, then p satisfies a Poincaré inequality with the same constant; hence the Poincaré constant is at most the log-Sobolev constant, $C_P \leqslant C_{LS}$. If $p \propto e^{-V}$ is α-strongly log-concave, that is, $V \succeq \alpha I_d$, then p satisfies a log-Sobolev inequality with constant $1/\alpha$.

We collect some properties of distributions satisfying LSI or PI.

Lemma 2.9.1 (Herbst, Sub-exponential and sub-gaussian concentration given log-Sobolev inequality, (Bakry et al., 2013, Pr. 5.4.1))**.** *Suppose that μ satisfies a log-Sobolev inequality with constant C_{LS}. Let f be a 1-Lipschitz function. Then*

1. *(Sub-exponential concentration) For any $t \in \mathbb{R}$,*

$$\mathbb{E}_\mu e^{tf} \leqslant e^{t\mathbb{E}_\mu f + \frac{C_{LS}t^2}{2}}.$$

2. *(Sub-gaussian concentration) For any* $t \in \left[0, \frac{1}{C_{LS}}\right)$,

$$\mathbb{E}_\mu e^{\frac{tf^2}{2}} \leqslant \frac{1}{\sqrt{1 - C_{LS}t}} \exp\left[\frac{t}{2(1 - C_{LS}t)}(\mathbb{E}_\mu f)^2\right].$$

Lemma 2.9.2 (Gaussian measure concentration for LSI, (Bakry et al., 2013, §5.4.2))**.** *Suppose that* μ *satisfies a log-Sobolev inequality with constant* C_{LS}*. Let* f *be a L-Lipschitz function. Then*

$$\mu\left(\left|f - \mathbb{E}_\mu f\right| \geqslant r\right) \leqslant 2e^{-\frac{r^2}{2C_{LS}L^2}}.$$

Lemma 2.9.3 ((Ge et al., 2018, Lemma G.10))**.** *Let* $V : \mathbb{R}^d \to \mathbb{R}$ *be a* α*-strongly convex and* β*-smooth function and let* P *be a probability measure with density function* $p(x) \propto e^{-V(x)}$*. Let* $x^* = \operatorname{argmin}_x V(x)$ *and* $\overline{x} = \mathbb{E}_P x$*. Then*

$$\|x^* - \overline{x}\| \leqslant \sqrt{\frac{d}{\alpha}}\left(\sqrt{\ln\left(\frac{\beta}{\alpha}\right)} + 5\right). \tag{2.34}$$

Theorem 2.9.4 ((Brascamp & Lieb, 2002, Theorem 5.1), (Hargé, 2004))**.** *Suppose the d-dimensional gaussian* $N(0, \Sigma)$ *has density* γ*. Let* $p = h \cdot \gamma$ *be a probability density.*

1. *If* h *is log-concave, and* g *is convex, then*

$$\int_{\mathbb{R}^d} g(x - \mathbb{E}_p x)p(x)\,dx \leqslant \int_{\mathbb{R}^d} g(x)\gamma(x)\,dx.$$

2. *If* h *is log-convex,*[1] *and* $g(x) = \langle x, y\rangle^\alpha$ *for some* $y \in \mathbb{R}^d$*,* $\alpha > 0$*, then*

$$\int_{\mathbb{R}^d} g(x - \mathbb{E}_p x)p(x)\,dx \geqslant \int_{\mathbb{R}^d} g(x)\gamma(x)\,dx.$$

Lemma 2.9.5. *Let* P *be a probability measure on* $\mathbb{R}^d$ *with density function* p *such that* $\ln p$ *is* C^1 *and* L*-smooth and* P *satisfies a Poincaré inequality with constant* C_P*. Then* $LC_P \geqslant 1$.

Proof. By the Poincaré inequality and Lemma 2.9.4(2), since p is equal to the density of $N(0, \frac{1}{L}I_d)$ multiplied by a log-convex function,

$$C_P \geqslant \mathbb{E}_P(x_1 - \mathbb{E}_P x_1)^2 \geqslant \mathbb{E}_{N(0,\frac{1}{L}I_d)} x_1^2 = \frac{1}{L}. \qquad \square$$

[1] Note that the sign is flipped in the theorem statement in (Brascamp & Lieb, 2002).

2.9.2 Lemmas on SMLD and DDPM

We give bounds on several quantities associated with the SMLD and DDPM processes at time t: the log-Sobolev constants (Lemma 2.9.7), the second moment (Lemma 2.9.8), and the warm start parameter (Lemma 2.9.9).

First, we note that for SMLD and DDPM, the conditional distribution of $\widetilde{x}_t$ given $\widetilde{x}_0$ is

$$\text{SMLD:} \qquad \widetilde{x}_t|\widetilde{x}_0 \sim N\left(x(0), \int_0^t g(s)^2\,ds \cdot I_d\right)$$

$$\text{DDPM:} \qquad \widetilde{x}_t|\widetilde{x}_0 \sim N\left(x(0)e^{-\frac{1}{2}\int_0^t g(s)^2\,ds}, (1-e^{-\int_0^t g(s)^2\,ds})I_d\right).$$

Hence

$$\widetilde{p}_t^{\text{SMLD}} = p_0 * N\left(0, \int_0^t g(s)^2\,ds \cdot I_d\right) \tag{2.35}$$

$$\widetilde{p}_t^{\text{DDPM}} = M_{e^{-\frac{1}{2}\int_0^t g(s)^2\,ds}\#} p_0 * N(0, (1-e^{-\int_0^t g(s)^2\,ds})I_d) \tag{2.36}$$

where M_c is multiplication by c.

Lemma 2.9.6 ((Chafaï, 2004)). *Let p, p' be two probability densities on $\mathbb{R}^d$. If p and p' satisfy log-Sobolev inequalities with constants C_{LS} and C'_{LS}, then $p * p'$ satisfies a log-Sobolev inequality with constant $C_{\text{LS}} + C'_{\text{LS}}$.*

Lemma 2.9.7 (Log-Sobolev constant for SMLD and DDPM). *Let $\widetilde{p}_t^{\text{SMLD}}$ and $\widetilde{p}_t^{\text{DDPM}}$ denote the distribution of the SMLD/DDPM processes at time t, when started at p_0. Let C_{LS} be the log-Sobolev constant of p_0. Then*

$$C_{\text{LS}}(\widetilde{p}_t^{\text{SMLD}}) \leqslant C_{\text{LS}} + \int_0^t g(s)^2\,ds$$

$$C_{\text{LS}}(\widetilde{p}_t^{\text{DDPM}}) \leqslant (C_{\text{LS}} - 1)e^{-\int_0^t g(s)^2\,ds} + 1 \leqslant \max\{C_{\text{LS}}, 1\}.$$

Note that the analogous statement for the Poincaré constant C_{P} holds for Lemma 2.9.6 and 2.9.7.

Proof. Note that if μ has log-Sobolev constant C_{LS} and T is a smooth L-Lipschitz map, then $T_\#\mu$ has log-Sobolev constant $\leqslant L^2 C_{\text{LS}}$. Applying Lemma 2.9.6 to (2.35) and (2.36) then finishes the proof. □

Lemma 2.9.8 (Second moment for SMLD and DDPM)**.** *Suppose that $\tilde{p}_0$ has finite second moment, then for $t \in [0, T]$:*

$$\mathbb{E}_{\tilde{p}_t}\left[\|x\|^2\right] = \mathbb{E}_{\tilde{p}_0}\left[\|x\|^2\right] + d\beta(t) \qquad \text{in SMLD,}$$

$$\mathbb{E}_{\tilde{p}_t}\left[\|x\|^2\right] = e^{-\beta(t)}\mathbb{E}_{\tilde{p}_0}\left[\|x\|^2\right] + d(1 - e^{-\beta(t)}) \leqslant \max\left\{\mathbb{E}_{\tilde{p}_0}\left[\|x\|^2\right], d\right\} \qquad \text{in DDPM,}$$

where $\beta(t) = \int_0^t g(s)^2 ds$.

Proof. Recall that in SMLD, $\tilde{x}_t \sim N(\tilde{x}_0, \beta(t) \cdot I_d)$. Let $y \sim N(0, \beta(t) \cdot I_d)$ be independent of $\tilde{x}_0$. Then

$$\mathbb{E}_{\tilde{p}_t}\left[\|x\|^2\right] = \mathbb{E}\left[\|\tilde{x}_0 + y\|^2\right] = \mathbb{E}\left[\|\tilde{x}_0\|^2\right] + \mathbb{E}\left[\|y\|^2\right] = \mathbb{E}\left[\|\tilde{x}_0\|^2\right] + d\beta(t).$$

In DDMP, $\tilde{x}_t \sim N(e^{-\frac{1}{2}\beta(t)}\tilde{x}_0, (1 - e^{-\beta(t)}) \cdot I_d)$. Choose $y \sim N(0, (1 - e^{-\beta(t)}) \cdot I_d)$ independent of $\tilde{x}_0$, then

$$\begin{aligned}\mathbb{E}_{\tilde{p}_t}\left[\|x\|^2\right] &= \mathbb{E}\left[\left\|e^{-\frac{1}{2}\beta(t)}\tilde{x}_0 + y\right\|^2\right] = \mathbb{E}\left[\left\|e^{-\frac{1}{2}\beta(t)}\tilde{x}_0\right\|^2\right] + \mathbb{E}\left[\|y\|^2\right] \\ &= e^{-\beta(t)}\mathbb{E}\left[\|\tilde{x}_0\|^2\right] + d(1 - e^{-\beta(t)}).\end{aligned}$$

□

Lemma 2.9.9 (Warm start for SMLD and DDPM)**.** *Suppose that p has log-Sobolev constant at most C_{LS} and $\left\|\mathbb{E}_{y\sim p} y\right\| \leqslant M_1$. Let φ_{σ^2} denote the density of $N(0, \sigma^2 I_d)$. Then for any σ^2,*

$$\chi^2(\varphi_{\sigma^2} || p * \varphi_{\sigma^2}) \leqslant 4\exp\left(\frac{d(2M_1 + 8C_{\mathrm{LS}})}{\sigma^2}\right)$$

Hence, letting $\sigma^2_{\mathrm{SMLD}} = \int_0^t g(s)^2\, ds$ and $\sigma^2_{\mathrm{DDPM}} = 1 - e^{-\int_0^t g(s)^2\, ds}$,

$$\chi^2(\varphi_{\sigma^2_{\mathrm{SMLD}}} || \widetilde{p}_t^{\mathrm{SMLD}}) \leqslant 4\exp\left(\frac{d(2M_1 + 8C_{\mathrm{LS}})}{\sigma^2_{\mathrm{SMLD}}}\right)$$

$$\chi^2(\varphi_{\sigma^2_{\mathrm{DDPM}}} || \widetilde{p}_t^{\mathrm{DDPM}}) \leqslant 4\exp\left(\frac{d\left(2e^{-\frac{1}{2}\int_0^t g(s)^2\, ds} M_1 + 8e^{-\int_0^t g(s)^2\, ds} C_{\mathrm{LS}}\right)}{\sigma^2_{\mathrm{DDPM}}}\right).$$

Proof. Let $R_x = (M_1 + 2\sqrt{C_{\mathrm{LS}}})\|x\|$. For a fixed x, note that $\mathbb{E}_{y\sim p}\langle y, x\rangle \leqslant \left\|\mathbb{E}_{y\sim p} y\right\| \|x\| \leqslant M_1 \|x\|$ by assumption. Then by Lemma 2.9.2,

$$\mathbb{P}(\langle y, x\rangle \geqslant R_x) \leqslant \mathbb{P}(|\langle y, x\rangle - \mathbb{E}_{y\sim p}\langle y, x\rangle| \geqslant 2\sqrt{C_{\mathrm{LS}}}\|x\|) \leqslant 2e^{-\frac{\left(2\sqrt{C_{\mathrm{LS}}}\|x\|\right)^2}{2C_{\mathrm{LS}}\|x\|^2}} \leqslant 2e^{-2} \leqslant \frac{1}{2}.$$

Hence

$$
\begin{aligned}
(p * \varphi_{\sigma^2})(x) &= \left(\frac{1}{2\pi\sigma^2}\right)^{d/2} \int_{\mathbb{R}^d} e^{-\frac{\|x+y\|^2}{2\sigma^2}} p(y)\, dy \\
&\geqslant \left(\frac{1}{2\pi\sigma^2}\right)^{d/2} e^{-\frac{\|x\|^2}{2\sigma^2}} \int_{\mathbb{R}^d} e^{-\frac{\langle x,y\rangle}{\sigma^2}} p(y)\, dy \\
&\geqslant \left(\frac{1}{2\pi\sigma^2}\right)^{d/2} e^{-\frac{\|x\|^2}{2\sigma^2}} \int_{\langle y,x\rangle \leqslant R_x} e^{-\frac{\langle x,y\rangle}{\sigma^2}} p(y)\, dy \\
&\geqslant \left(\frac{1}{2\pi\sigma^2}\right)^{d/2} e^{-\frac{\|x\|^2}{2\sigma^2}} \int_{\langle y,x\rangle \leqslant R_x} e^{-(M_1\|x\| + 2\sqrt{C_{LS}}\|x\|)/\sigma^2} p(y)\, dy \\
&\geqslant \left(\frac{1}{2\pi\sigma^2}\right)^{d/2} e^{-\frac{\|x\|^2}{2\sigma^2}} e^{-\frac{\|x\|^2}{8\sigma^2 d} - \frac{2M_1^2 d}{\sigma^2} - \frac{\|x\|^2}{8\sigma^2 d} - \frac{8C_{LS} d}{\sigma^2}} \int_{\langle y,x\rangle \leqslant R_x} p(y)\, dy \\
&\geqslant \left(\frac{1}{2\pi\sigma^2}\right)^{d/2} e^{-\frac{\|x\|^2}{2\sigma^2\left(1-\frac{1}{2d}\right)^{-1}}} e^{-\frac{d(8C_{LS}+2M_1^2)}{\sigma^2}} \cdot \frac{1}{2} \\
&\geqslant \frac{1}{2} e^{-\frac{d(8C_{LS}+2M_1^2)}{\sigma^2}} \left(1-\frac{1}{2d}\right)^{d/2} \varphi_{\frac{\sigma^2}{1-\frac{1}{2d}}}.
\end{aligned}
$$

Using the fact that $\chi^2(N(0,\Sigma_2)||N(0,\Sigma_1)) = \frac{|\Sigma_1|^{1/2}}{|\Sigma_2|}|(2\Sigma_2^{-1} - \Sigma_1^{-1})|^{-\frac{1}{2}} - 1$, we have

$$
\begin{aligned}
\chi^2(\varphi_{\sigma^2}||p * \varphi_{\sigma^2}) + 1 &\leqslant 2 \cdot e^{\frac{d(8C_{LS}+2M_1^2)}{\sigma^2}} \left(1-\frac{1}{2d}\right)^{-\frac{d}{2}} \left[\chi^2\left(\varphi_{\sigma^2}||\varphi_{\frac{\sigma^2}{1-\frac{1}{2d}}}\right) + 1\right] \\
&= 2 \cdot e^{\frac{d(8C_{LS}+2M_1^2)}{\sigma^2}} \left(1-\frac{1}{2d}\right)^{-\frac{d}{2}} \left(1-\frac{1}{2d}\right)^{-\frac{d}{2}} \left(2 - \left(1-\frac{1}{2d}\right)\right)^{-\frac{d}{2}} \\
&\leqslant 2 \cdot e^{\frac{d(8C_{LS}+2M_1^2)}{\sigma^2}} \left(1-\frac{1}{2d}\right)^{-d} \leqslant 4e^{\frac{d(8C_{LS}+2M_1^2)}{\sigma^2}}
\end{aligned}
$$

The corollary inequalities then follow from (2.35) and (2.36), where for DDPM, we use the fact that $M_{e^{-\frac{1}{2}\int_0^t g(s)^2\, ds}\#} p_0$ has mean $e^{-\frac{1}{2}\int_0^t g(s)^2\, ds} \cdot \mathbb{E}_p x$ and LSI-constant $(e^{-\frac{1}{2}\int_0^t g(s)^2\, ds})^2 C_{LS}$. $\square$

3. Convergence of SGMs for general data distributions

3.1 Background and Preliminaries

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 (Dathathri et al., 2019; Grathwohl et al., 2019; Jing et al., 2022; Meng et al., 2021; Song, Durkan, et al., 2021; Song & Ermon, 2019, 2020; Song, Shen, et al., 2021; Song, Sohl-Dickstein, et al., 2020), including being a key component of text-to-image systems (Ramesh et al., 2022), 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 Bortoli, 2022; De Bortoli et al., 2021; Lee et al., 2022), 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. For example, considering the hallmark application of generating images from text, we expect the distribution of images to be (a) multimodal, and hence not satisfying functional inequalities with reasonable constants, and (b) supported on lower-dimensional manifolds, and hence not smooth. However, SGM still performs remarkably well in these settings. Indeed, this is one relative advantage to other approaches to generative modeling such as generative adversarial networks, which can struggle to learn multimodal distributions (Arora et al., 2018).

In this chapter, we aim to develop theoretical convergence guarantees with polynomial complexity for SGM under minimal data assumptions.

3.1.1 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 (Song, Sohl-Dickstein, et al., 2020)

$$d\widetilde{x}_t = f(\widetilde{x}_t, t)\, dt + g(t)\, dw_t, \quad t \in [0, T], \tag{3.1}$$

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,

$$d\widetilde{x}_t = [f(\widetilde{x}_t, t) - g(t)^2 \nabla \ln \widetilde{p}_t(\widetilde{x}_t)]\, dt + g(t)\, d\widetilde{w}_t, \quad t \in [0, T], \tag{3.2}$$

where $\widetilde{w}_t$ is a backward Brownian motion (Anderson, 1982). Because the forward process transforms the data distribution to noise, the hope is to use the backwards process to transform noise into samples.

In practice, when we only have sample access to p_{data}, the score function $\nabla \ln \widetilde{p}_t$ is not available. A key mechanism behind SGM is that the score function is learnable from data, through empirically minimizing a de-noising objective evaluated at noisy samples $\widetilde{x}_t$ (Vincent, 2011). The samples $\widetilde{x}_t$ are obtained by evolving the forward SDE starting from the data samples $\widetilde{x}_0$, and the optimization is done within an expressive function class such as neural networks. If the score function is successfully approximated in this way, then the L^2-error $\mathbb{E}_{\widetilde{p}_t}[\|\nabla \ln \widetilde{p}_t(x) - s(x,t)\|^2]$ will be small. The natural question is then the following:

> Given L^2-error bounds of the score function, how close is the distribution generated by (3.2) (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. Indeed, prior work on Langevin Monte Carlo (Erdogdu et al., 2021) and related sampling algorithms only apply when the score function is known exactly, or with suitable modification, known up to L^∞-error. L^2-error has a genuinely different effect

from L^∞-error, as it can cause the stationary distribution for Langevin Monte Carlo to be arbitrarily diffferent (Lee et al., 2022), necessitating a "medium-time" analysis.

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.

3.1.2 Prior work on convergence guarantees

We highlight two recent works which make progress on this problem. (Lee et al., 2022) 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 Bortoli, 2022) instead make minimal data assumptions, giving convergence in Wasserstein distance for distributions with bounded support $\mathcal{M}$. In particular, this covers the case of distributions supported on lower-dimensional manifolds, where guarantees in TV distance are unattainable. However, for general distributions, their guarantees have exponential dependence on the diameter of $\mathcal{M}$ and the inverse of the desired error ($\exp(O(\operatorname{diam}(\mathcal{M})^2/\varepsilon))$), and for smooth distributions, an improved, but still exponential dependence on the growth rate of the Hessian $\nabla^2 \ln \widetilde{p}_t$ as the noise approaches 0 ($\exp(\widetilde{O}(\Gamma))$ for distributions with $\|\nabla^2 \ln \widetilde{p}_t\| \leqslant \Gamma/\sigma_t^2$).

We note that other works also analyze the generalization error of a learned score estimate (Block et al., 2020; De Bortoli, 2022). 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.

3.1.3 Our contributions

In this chapter, we analyze convergence in the most general setting of distributions of bounded support, as in (De Bortoli, 2022). 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. This gives theoretical grounding to the empirical success of SGM on data distributions that are often multimodal and non-smooth.

We streamline the χ^2-based analysis of (Lee et al., 2022), 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), and then a structural result that the distributions arising from the diffusion process can be slightly modified as to satisfy the desired inequality, through decomposition into distributions that do satisfy a log-Sobolev inequality.

3.2 Main results

To state our results, we will consider a specific type of SGM called denoising diffusion probabilistic modeling (DDPM) (Ho et al., 2020), where in the forward SDE (3.1), $f(x,t) = -\frac{1}{2}g(t)^2 x$ 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

$$m_t \widetilde{x}_0 + \sigma_t z, \text{ where}$$

$$m_t = \exp\left[-\frac{1}{2}\int_0^t g(s)^2\,ds\right],\ \sigma_t^2 = 1 - \exp\left[-\int_0^t g(s)^2\,ds\right],\ \text{and } z \sim N(0,I). \tag{3.3}$$

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

$$dx_t = \frac{1}{2}g(T-t)^2\left(x_t + 2\nabla \ln p_t(x_t)\right)\,dt + g(T-t)\,dw_t \tag{3.4}$$

with the exponential integrator discretization (Zhang & Chen, 2022). Denoting this discretized process by z_t, and letting $h_k = t_{k+1} - t_k$ and $\eta_{k+1} \sim N(0, I_d)$.

$$z_{t_{k+1}} = z_{t_k} + \gamma_{1,k}(z_{t_k} + 2s(T - t_k, z_{t_k})) + \sqrt{\gamma_{2,k}} \cdot \eta_{k+1}, \tag{3.5}$$

$$\text{where } \gamma_{1,k} = \exp\left[\frac{1}{2}G_{t_k,t_{k+1}}\right] - 1,\ \gamma_{2,k} = \exp\left[G_{t_k,t_{k+1}}\right] - 1, \text{ and } G_{t',t} := \int_{t'}^{t} g(T-s)^2\, ds. \tag{3.6}$$

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

$$z_0 \sim q_0 = p_{\text{prior}} := N(0, \sigma_T^2 I_d) \approx N(0, I_d). \tag{3.7}$$

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 Lee et al., 2022, §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 $0 = t_0 < t_1 < \cdots < t_K \leqslant T$ has been chosen.

Assumption 3 (L^2 score error)**.** *For any $t \in \{T - t_0, \ldots, T - t_K\}$, the error in the score estimate is bounded in $L^2(\widetilde{p}_t)$:*

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

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 4 (Bounded support)**.** p_{data} *is supported on* $B_R(0) := \{x \in \mathbb{R}^d : \|x\| \leqslant R\}$.

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 4 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 Bortoli, 2022 and will give better dependencies. Note that De Bortoli, 2022, Theorem I.8 shows a (nonuniform) version of Assumption 5 holds when p_0 is a smooth density on a convex submanifold.

Assumption 5. *The following bound of the Hessian of the log-pdf holds for any $t > 0$ and x:*

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

for some constant $C > 0$.

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

Assumption 6. *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 3 DDPM with exponential integrator (Song, Sohl-Dickstein, et al., 2020; Zhang & Chen, 2022)

INPUT: Time T; discretization points $0 = t_0 < t_1 < \cdots < t_N \leqslant T$; 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 (3.7).
for k from 1 to N **do**
  Compute z_{t_k} from $z_{t_{k-1}}$ using (3.5).
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 3.2.1 (Wasserstein+TV error for distributions with bounded support)**.** *Suppose that Assumption 3 and 4 hold with $R \geqslant \sqrt{d}$. 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 5 holds with $C \geqslant R^2$, 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 (3.2), 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 3.7.2. We note that under the Hessian bound (Assumption 5), 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 3.2.2 (Wasserstein error for distributions with bounded support)**.** *In the same setting as Theorem 3.2.1, consider the distribution $\widehat{q}_{t_N}$ of the scaled and truncated output $\widehat{x}_{t_N}$ of DDPM. If Assumptions 3 and 4 hold with $R \geqslant \sqrt{d}$ and $\varepsilon_\sigma = \widetilde{O}\left(\frac{\varepsilon_{\text{W}}^{18}}{R^{22} d^{4.75}}\right)$, then with appropriate (polynomial) choice of parameters, $W_2(\widehat{q}_{t_K}, p_{\text{data}}) \leqslant \varepsilon_{\text{W}}$. If in addition Assumption 5 holds with $C \geqslant R^2$, then $\varepsilon_\sigma = \widetilde{O}\left(\frac{\varepsilon_{\text{W}}^8}{C^2 R^8 d}\right)$ suffices.*

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

Theorem 3.2.3 (TV error for distributions under smoothness assumption)**.** *Suppose that Assumptions 3 and 6 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* $\mathrm{TV}(q_{t_N}, p_{\mathrm{data}}) \leqslant \varepsilon_{\mathrm{TV}}$. *If in addition Assumption 5 holds with* $C \geqslant R^2$, *then* $\varepsilon_\sigma = \widetilde{O}\left(\frac{\varepsilon_{\mathrm{TV}}^4}{C^2 d}\right)$ *suffices.*

A more precise statement can be found as Theorem 3.7.3, which also works more generally with sufficient tail decay. We note that this result can be derived directly by combining Theorem 3.7.2 and a TV error bound between P_{data} and p_{t_N} (Lemma 3.6.4) depending on the smoothness of P_{data}.

3.3 Notation and proof overview

We let $\widetilde{p}_t$ denote the density of $\widetilde{x}_t$ under the forward process (3.1). 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 (3.3),

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

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 $\sigma_t^2 \leqslant \min\{1, t\}$ and $\sigma_t^2 = \Theta(\min\{1, t\})$.

We will let z_t denote the (interpolated) discrete process (see (3.13)) and let q_t be the density of z_t. We define

$$\phi_t(x) = \frac{q_t(x)}{p_t(x)}, \qquad \psi_t(x) = \frac{\phi_t(x)}{\mathbb{E}_{p_t}\phi_t^2}, \tag{3.8}$$

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 (3.6).

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

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

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

$$\mathscr{E}_p(f,g) := \int_{\mathbb{R}^d} \langle \nabla f, \nabla g\rangle \, P(dx) = \int_{\mathbb{R}^d} \langle \nabla f, \nabla g\rangle \, p(x)\, dx; \tag{3.9}$$

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,

$$\mathrm{KL}(q||p) \leqslant \frac{C_{\mathrm{LS}}}{2}\mathscr{E}_p\left(\frac{q}{p}, \ln\frac{q}{p}\right) = \frac{C_{\mathrm{LS}}}{2}\int_{\mathbb{R}^d}\left\|\nabla \ln\frac{q(x)}{p(x)}\right\|^2 q(x)\, dx. \tag{3.10}$$

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 $f \geqslant 0$,

$$\mathrm{Ent}_p(f) \leqslant \frac{C_{\mathrm{LS}}}{2}\mathscr{E}_p\left(f, \ln f\right) = \frac{C_{\mathrm{LS}}}{2}\int_{\mathbb{R}^d}\|\nabla \ln f(x)\|^2 f(x)p(x)dx. \tag{3.11}$$

3.3.1 Proof overview

Our proof uses the framework by Lee et al., 2022 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

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

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 3.4.10), 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 Lee et al., 2022 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 3.5.2, 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 3.6, 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 3.6.5 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 3.7). 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 3.8 we show that the Hessian is always bounded by $O\left(\frac{d}{\sigma_t^2}\right)$ with high probability (cf. Assumption 5). We speculate that a high-probability rather than uniform bound on the Hessian (as in Lemma 3.4.13) can be used to obtain better dependencies, and leave this as an open problem.

3.4 DDPM with L^∞-accurate score estimate

We consider the error between the exact backwards SDE (3.4) and the exponential integrator with estimated score (3.5). In this section, we bound the error assuming that the score estimate s is accurate in L^∞.

Assumption 7 (L^∞ score error)**.** *For any $t \in \{T - t_0, \ldots, T - t_K\}$, the error in the score estimate*

is bounded:

$$\|\nabla \ln \widetilde{p}_t - s(\cdot, t)\|_\infty = \sup_{x \in \mathbb{R}^d} \|\nabla \ln \widetilde{p}_t(x) - s(x,t)\| \leqslant \varepsilon_{\infty,t}^2 \tag{3.12}$$

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

In Section 3.7, 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 (3.5), for $t \in [t_k, t_{k+1}]$:

$$dz_t = g(T-t)^2 \left(\frac{1}{2} z_t + s(z_{t_k}, T - t_k) \right) dt + g(T-t)\, dw_t. \tag{3.13}$$

Integration gives that

$$\begin{aligned} z_t - z_{t_k} = &\left(\exp\left(\frac{1}{2} G_{t_k,t} \right) - 1 \right) (z_{t_k} + 2s(z_{t_k}, T - t_k)) \\ &+ \int_{t_k}^t \exp\left(\frac{1}{2} \int_{t_k}^{t'} g(T-t'')^2\, dt'' \right) g(t')\, dw_{t'}, \end{aligned} \tag{3.14}$$

where $G_{t',t}$ is defined in (3.6).

Letting q_t be the distribution of z_t and p_t be the distribution of x_t, we have by (Lee et al., 2022, Lemma A.2) that

$$\begin{aligned} \frac{d}{dt} \chi^2(q_t || p_t) = &-g(T-t)^2 \mathscr{E}_{p_t}\left(\frac{q_t}{p_t} \right) \\ &+ 2\mathbb{E}\left[\left\langle g(T-t)^2 (s(z_{t_k}, T-t_k) - \nabla \ln p_t(z_t)), \nabla \frac{q_t(z_t)}{p_t(z_t)} \right\rangle \right]. \end{aligned} \tag{3.15}$$

(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 (3.8): $\phi_t(x) = \frac{q_t(x)}{p_t(x)}$, $\psi_t(x) = \frac{\phi_t(x)}{\mathbb{E}_{p_t} \phi_t^2}$.

To bound (3.15), we use the following lemma.

Lemma 3.4.1 (cf. Erdogdu et al., 2021, Lemma 1, Lee et al., 2022, Lemma A.3)**.** *For any* $C > 0$ *and any* $\mathbb{R}^d$*-valued random variable* u*, we have*

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

Proof. By Young's inequality,

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

Lemma 3.4.2. *Suppose that* (3.12) *holds for* $t = T - t_k$, $\nabla \ln p_{t_k}(x)$ *is* L_{T-t_k}*-Lipschitz,* g *is non-decreasing, and that* $h_k \leqslant \frac{1}{20L_{T-t_k}g(T-t_k)^2}$. *Then we have for* $t \in [t_k, t_{k+1}]$ *that*

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

Proof. We bound the second term on the RHS of (3.15). By Lemma 3.4.1,

$$\begin{aligned}
&\mathbb{E}\left[\left\langle s(z_{t_k}, T-t_k) - \nabla \ln p_t(z_t), \nabla \frac{q_t(z_t)}{p_t(z_t)}\right\rangle\right] \\
&\leqslant (\chi^2(q_t||p_t)+1)\mathbb{E}\left[\|s(z_{t_k}, T-t_k) - \nabla \ln p_t(z_t)\|^2 \psi_t(z_t)\right] + \frac{1}{4}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right). \qquad (3.16)
\end{aligned}$$

Now

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

and

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

by definition of $\varepsilon_{\infty,t}$, so by Lemma 3.4.3,

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

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

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

Substituting into (3.16) and that inequality into (3.15) give the conclusion. □

Lemma 3.4.3. *Suppose that* $h_k \leqslant \frac{1}{2g(T-t_k)^2}$. *Then for* $t \in [t_{k+1}, t_k]$,

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

Proof. Consider (3.14). The assumption on h_k implies $G_{t_k,t} \leqslant G_{t_k,t_{k+1}} \leqslant \frac{1}{2}$, so $\exp\left(\frac{1}{2}G_{t_k,t}\right) - 1 \leqslant G_{t_k,t}$. Let Y denote the last term of (3.14). Then

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

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

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

and

$$\mathbb{E}\left[\|z_t - z_{t_k}\|^2 \psi_t(z_t)\right] \leqslant 16G_{t_k,t}^2\Big[\mathbb{E}[\psi_t(z_t)\|z_t\|^2] + 4\mathbb{E}[\psi_t(z_t)\|s(z_{t_k}, T - t_k) - \nabla \ln p_t(z_t)\|^2] + 16\mathbb{E}[\psi_t(z_t)\|\nabla \ln p_t(x_t)\|^2]\Big] + 16\mathbb{E}[\psi_t(z_t)\|Y\|^2].$$

By Lemma 3.4.4,

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

The lemma follows. □

Lemma 3.4.4. *For $t \in [t_k, t_{k+1}]$,*

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

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 (3.13) does agree with the discrete-time process (3.5) at $t = t_{k+1}$.) Using the Donsker-Varadhan variational principle, for any random variable X,

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

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 $\mathrm{d}\tilde{\mathbb{P}}\mathbb{P}(z_t) = \psi_t(z_t)$, we can bound

$$\begin{aligned}
\tilde{\mathbb{E}} \|Y\|^2 &\leqslant 2\mathbb{E}\left[\|Y\|^2\right] + 2\tilde{\mathbb{E}}\left[(Y - \mathbb{E}\|Y\|)^2\right] \\
&\leqslant 2\mathbb{E}\left[\|Y\|^2\right] + \frac{2}{c}\left[\mathrm{KL}(\tilde{\mathbb{P}}||\mathbb{P}) + \ln \mathbb{E} \exp\left(c\left(\|Y\| - \mathbb{E}\|Y\|\right)^2\right)\right] \qquad (3.17) \\
&\leqslant 2d(\exp(G_{t_k,t}) - 1) + \frac{2}{c}\left[\mathrm{KL}(\tilde{\mathbb{P}}||\mathbb{P}) + \ln \mathbb{E} \exp\left(c\left(\|Y\| - \mathbb{E}\|Y\|\right)^2\right)\right]. \qquad (3.18)
\end{aligned}$$

Now following Chewi et al., 2021, Theorem 4, we set $c = \frac{1}{8(\exp(G_{t_k,t})-1)}$, so that

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

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

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

Substituting everything into (3.18) gives the desired inequality. □

Let

$$K_z = \mathbb{E}\left[\psi_t(z_t)\|z_t\|^2\right] \qquad (3.19)$$

$$K_V = \mathbb{E}\left[\psi_t(z_t)\|\nabla \ln p_t(z_t)\|^2\right] \qquad (3.20)$$

$$K_{\Delta V} = \mathbb{E}\left[\psi_t(z_t)\|\nabla \ln p_{t_k}(z_t) - \nabla \ln p_t(z_t)\|^2\right] \qquad (3.21)$$

$$K = \mathrm{KL}(\psi_t q_t||p_t). \qquad (3.22)$$

In order to bound the RHS in Lemma 3.4.2, we need to bound all four of these quantities, which we do in Lemma 3.4.5, 3.4.6, 3.4.8, and Section 3.5, respectively. The main innovation

in our analysis compared to Lee et al., 2022 is a new way to bound K, which we present in a separate section. First we bound K_z. Recall the norm

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

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

Lemma 3.4.5. *For $t \in [t_k, t_{k+1}]$,*

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

Proof. By the Donsker-Varadhan variational principle,

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

for any $s > 0$. Choosing $s = 2\|x_t\|_{2,\psi_2}^{-2}$, we have $\mathbb{E}_{p_t}\left[e^{\frac{s}{2}\|x\|^2}\right] \leqslant 2$, 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 3.4.6 (Lee et al., 2022, Corollary C.7, Chewi et al., 2021, Lemma 16)**.**

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

We use the following lemma to bound $K_{\Delta V}$ in Lemma 3.4.8.

Lemma 3.4.7 (Lee et al., 2022, 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 $\sigma^2 \leqslant \frac{1}{2\alpha^2 L}$,*

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

Lemma 3.4.8. *Suppose that* $h_k \leqslant \frac{1}{4Lg(T-t_k)^2}$ *where* $\nabla \ln p_t$ *is* L_{T-t}*-smooth (*$L_{T-t} \geqslant 1$*) and* $L = \max_{t\in[t_k,t_{k+1}]} L_{T-t}$*. For* $t \in [t_k, t_{k+1}]$*,*

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

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 $h_k \leqslant \frac{1}{4g(T-t_k)^2}$,

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

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

We note that

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

so the hypothesis of Lemma 3.4.7 is satisfied. Using Lemma 3.4.7, we obtain

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

□

Now we put everything together. Write $G_t = G_{t_k,t}$ for short. Suppose L_t is non-increasing. By Lemma 3.4.2,

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

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

By Lemma 3.4.8, $K_{\Delta V} \leqslant 25L_{T-t}^2(8G_td + G_t^2K_z) + 100L_{T-t}^2G_t^2K_V$, so

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

By Lemma 3.4.5, $K_z \leqslant \|x_t\|_{2,\psi_2}^2(K+\ln 2)$, and by Corollary 3.4.6, $K_V \leqslant \frac{4}{\chi^2(q_t||p_t)+1}\cdot\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 2dL$, so

$$E \leqslant 41G_t^2L_{T-t}^2\left(\|x_t\|_{2,\psi_2}^2(K+\ln 2)\right) + 356G_t^2L_{T-t}^2\left(\frac{4}{\chi^2(q_t||p_t)+1}\cdot\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 2dL\right)$$

$$+ 64G_tL_{T-t}^2(8K+6d+16\ln 2) + \varepsilon_{\infty,T-t_k}^2.$$

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

$$E \leqslant {\varepsilon'_{h_k}}^2\left[\|x_t\|_{2,\psi_2}^2(K+\ln 2) + \left(\frac{4}{\chi^2(q_t||p_t)+1}\cdot\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + 2dL_{T-t}\right)\right]$$

$$+ 4\varepsilon'_{h_k}L_{T-t}(8K+2d+16\ln 2) + \varepsilon_{\infty,T-t_k}^2.$$

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

$$12g(T-t)^2(\chi^2(q_t||p_t)+1)\cdot E$$

$$\leqslant 12g(T-t)^2\Big[\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right)({\varepsilon'_{h_k}}^2\cdot(A_{T-t}M_{T-t}+4) + \varepsilon'_{h_k}\cdot 32L_{T-t}A_{T-t})$$

$$+ (\chi^2(q_t||p_t)+1)({\varepsilon'_{h_k}}^2\cdot((B_{T-t}+\ln 2)M_{T-t}+2dL)$$

$$+ \varepsilon'_{h_k}\cdot L_{T-t}(8B_{T-t}+6d+16\ln 2)) + \varepsilon_{\infty,T-t_k}^2\Big].$$

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

$$\frac{d}{dt}\chi^2(q_t||p_t) \leqslant 12g(T-t)^2\Big[(\chi^2(q_t||p_t)+1)({\varepsilon'_{h_k}}^2\cdot((B_{T-t}+\ln 2)M_{T-t}+2dL_{T-t})$$

$$+ \varepsilon'_{h_k}\cdot L_{T-t}(8B_{T-t}+6d+16\ln 2)) + \varepsilon_{\infty,T-t_k}^2\Big].$$

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

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

Integration gives

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

Taking $\varepsilon' = \frac{\varepsilon}{\ln\left(\frac{T}{T-t_N}\right)}$ then gives the following Theorem 3.4.10. We first introduce a technical assumption.

Definition 3.4.9. *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 3.4.10. *Suppose that the following hold.*

1. *Assumption 7 holds for $\varepsilon_{\infty,t}$.*
2. *$\|\widetilde{x}_t\|^2_{2,\psi_2} \leqslant M_t$.*
3. *The KL bound $\mathrm{KL}(\psi_t q_t||p_t) \leqslant \frac{A_{T-t}}{\chi^2(q_t||p_t)+1} + B_{T-t}$ 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

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

$$\text{where } \varepsilon'_{h_k} = \min\left\{\frac{1}{\sqrt{A_{T-t_k}M_{T-t_k}+1}}, \frac{1}{L_{T-t_k}A_{T-t_k}}, \right.$$

$$\frac{\sqrt{\varepsilon/\ln\left(\frac{T}{T-t_N}\right)}}{g(T-t_k)\sqrt{(T-t_k)((B_{T-t_k}+1)M_{T-t_k}+dL_{T-t_k})}},$$

$$\left.\frac{\varepsilon/\ln\left(\frac{T}{T-t_N}\right)}{g(T-t_k)^2(T-t_k)L_{T-t_k}(B_{T-t_k}+d)}\right\},$$

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

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

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 $h_k = t_{k+1} - t_k \leqslant \frac{T-t_k}{2}$. □

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 3.31, Lemma 3.5.2 will give the existence of a distribution that does, and is close in TV-error. Later in Section 3.6, we show that this will have a small effect on the score function, and hence allow us to prove our main theorems.

Corollary 3.4.11. *Suppose that Assumptions 7 and 4 hold, $R^2 \geqslant d$, $g \equiv 1$, and that $\widetilde{P}_0$ is such that the KL inequality* (3.31) *holds. Let $\delta = T - t_N$. If $0 < \delta, \varepsilon < \frac{1}{2}$ and*

$$h_k = O\left(\frac{\varepsilon}{\max\{T-t_k, (T-t_k)^{-3}\}R^4 d \ln\left(\frac{T}{\delta}\right)\ln\left(\frac{R}{\delta\varepsilon_K}\right)}\right),$$

then for any $0 \leqslant k \leqslant N$,

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

Proof. For $g \equiv 1$, note that $\sigma^2_{T-t} = \Theta(\min\{T-t, 1\})$. From Lemma 3.4.13, we can choose

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

From Lemma 3.4.15, we can choose

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

The KL inequality (3.31) 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

$$\varepsilon'_{h_k} = O\left(\frac{1}{\sqrt{A_{T-t_k}M_{T-t_k}+1}}\right) \qquad \Longleftarrow \varepsilon'_{h_k} = O\left(\frac{1}{\max\{R, \sqrt{d}\}}\right) \tag{3.23}$$

$$\varepsilon'_{h_k} = O\left(\frac{1}{L_{T-t_k}A_{T-t_k}}\right) \qquad \Longleftarrow \varepsilon'_{h_k} = O\left(\frac{T-t_k}{R^2}\right) \tag{3.24}$$

$$\varepsilon'_{h_k} = O\left(\frac{\sqrt{\varepsilon/\ln\left(\frac{T}{\delta}\right)}}{\sqrt{(T-t_k)((B_{T-t_k}+1)M_{T-t_k} + dL_{T-t_k})}}\right). \tag{3.25}$$

For $T - t_k \leqslant 1$, (3.25) 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 3.4.10 gives the result. □

Above, we use the Hessian bound $\left\|\nabla^2 \ln p_t(x)\right\| \leqslant \frac{R^2}{\sigma_t^4}$ given in Lemma 3.4.13. Under the stronger smoothness assumption given by Assumption 5, we can take the step sizes to be larger.

Corollary 3.4.12. *Suppose that Assumptions 7, 4, 5 hold, $C \geqslant R^2 \geqslant d$, $g \equiv 1$, and that $\widetilde{P}_0$ is such that the KL inequality* (3.31) *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 $0 \leqslant k \leqslant N$,*

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

Proof. We instead have the bound $L_t = \frac{C}{\sigma_t^2}$. The requirement (3.23) stays the same, while (3.24) is implied by $\varepsilon'_{h_k} = O(1/C)$. Inequality (3.25), for $T - t_k \leqslant 1$, 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 $C \geqslant R^2$, $\varepsilon \leqslant 1/\sqrt{T}$, implies all the others. □

3.4.1 Auxiliary bounds

In this section we give bounds on the Hessian (L_t, Lemma 3.4.13), initial χ^2 divergence $\chi^2(q_0||p_0)$ (Lemma 3.4.14), and Orlicz norm (M_t, Lemma 3.4.15).

Lemma 3.4.13 (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 $N(0, \sigma^2 I_d)$,*

$$\left\|\nabla^2 \ln(\mu * \varphi_{\sigma^2}(x))\right\| \leqslant \max\left\{\frac{R^2}{\sigma^4}, \frac{1}{\sigma^2}\right\}. \tag{3.26}$$

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

$$\left\|\nabla^2 \ln \widetilde{p}_t(x)\right\| \leqslant \frac{R^2}{\sigma_t^4}. \tag{3.27}$$

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:

$$\nabla \ln(\mu * \varphi_{\sigma^2}(x)) = \frac{\nabla \int_{\mathbb{R}^d} e^{-\frac{\|x-u\|^2}{2\sigma^2}}\mu(du)}{\int_{\mathbb{R}^d} e^{-\frac{\|x-u\|^2}{2\sigma^2}}\mu(du)} = \frac{\int_{\mathbb{R}^d} -\frac{x-u}{\sigma^2} e^{-\frac{\|x-u\|^2}{2\sigma^2}}\mu(du)}{\int_{\mathbb{R}^d} e^{-\frac{\|x-u\|^2}{2\sigma^2}}\mu(du)} = -\frac{1}{\sigma^2}\mathbb{E}_{\mu_{x,\sigma^2}}(x - u) \tag{3.28}$$

$$\nabla^2 \ln(\mu * \varphi_{\sigma^2}(x)) = \frac{1}{\sigma^4}\operatorname{Cov}_{\mu_{x,\sigma^2}}(x - u) - \frac{1}{\sigma^2} I_d = \frac{1}{\sigma^4}\operatorname{Cov}_{\mu_{x,\sigma^2}}(u) - \frac{1}{\sigma^2} I_d. \tag{3.29}$$

The covariance of a distribution supported on a set of radius R is bounded by R^2 in operator norm. Inequality (3.26) then follows from (3.29).

For (3.27), note that $\widetilde{p}_t = M_{m_t\sharp}\widetilde{P}_0 * \varphi_{\sigma_t^2}$, where m_t is given by (3.3) 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 $\sigma_t \leqslant 1$, the result follows. □

Lemma 3.4.14 (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*

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

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

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

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

$$\leqslant \exp\left[\frac{\|x_0\|^2\exp(-G_{0,T})}{1-\exp(-G_{0,T})}\right] - 1 \leqslant \exp\left(\frac{R^2\exp(-G_{0,T})}{1-\exp(-G_{0,T})}\right) - 1$$

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

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

Lemma 3.4.15 (Subgaussian bound)**.** *Suppose $\widetilde{P}_0$ is supported on $B_R(0)$. Then for $X \sim \widetilde{p}_t$,*

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

where m_t, σ_t are as in (3.3) *and C_1 is an absolute constant.*

Proof. Let $Y \sim \widetilde{P}_0$ s.t. $X = m_t Y + \sigma_t \xi$ 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 $p \geqslant 1$,

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

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

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

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

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

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

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

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 $1/(1-x) \leqslant e^{2x}$ which is valid for $x \in [0, 1/2]$. It follows that

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

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

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

which implies that

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

3.5 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 (3.8). 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 3.5.1. *Suppose that $\widetilde{P}_0$ is a probability measure on $\mathbb{R}^d$ such that*

$$\widetilde{P}_0 = \sum_{j=1}^{m} w_j \widetilde{P}_{j,0}, \tag{3.30}$$

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 (3.1) *for time t, and $p_t = \widetilde{p}_{T-t}$, $p_{j,t} = \widetilde{p}_{j,T-t}$. Let $w_{\min} = \min_{1 \leqslant j \leqslant m} w_j$ 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* (3.8)

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

While we need p_t to satisfy a log-Sobolev inequality to get a bound of the form $\frac{C}{\chi^2(q_t||p_t)+1}\mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right)$ (Lee et al., 2022, 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 3.5.2 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},

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

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

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

Lemma 3.5.2. *Suppose $0 < \varepsilon_K < \frac{1}{2}$, and that $\overline{P}_0$ is a probability measure such that $\overline{P}_0(\mathcal{M}) \geqslant 1 - \frac{\varepsilon_K}{8}$. 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 $\chi^2(\widetilde{P}_0||\overline{P}_0) \leqslant \varepsilon_K$ and considering the DDPM process started with $\widetilde{P}_0$, for all $0 \leqslant t \leqslant T - \delta$,*

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

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

$$\mathrm{KL}(\psi_t q_t || p_t) \leqslant \left(\frac{6(1+e)\sigma_{T-t}^2}{\chi^2(q_t||p_t)+1} \mathscr{E}_{p_t}\left(\frac{q_t}{p_t}\right) + \ln\left(\frac{1}{\varepsilon_K}\right) + d\ln\left(1 + \frac{4R}{\sigma_\delta}\right)\right). \tag{3.31}$$

Proof. Partition $\mathcal{M}$ into disjoint subsets $\mathcal{M}_j$, $1 \leqslant j \leqslant N := \mathcal{N}(\mathcal{M}, \sigma_\delta/2)$ 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 $Z = \sum_{j:w_j \geqslant \frac{\varepsilon_K}{8N}} w_j$ and

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

and define

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

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

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

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 $m_t\sigma \leqslant \sigma$. By Theorem 1 of H.-B. Chen et al., 2021, noting that

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

when $\Sigma = \sigma^2 I$ and $\|\mu_2 - \mu_1\| \leqslant \sigma$, $\widetilde{P}_{j,t} = (M_{m_t\sharp}\widetilde{P}_{j,0}) * \varphi_{\sigma^2}$ satisfies a log-Sobolev inequality with constant $6(1+e)\sigma_t^2$. The result then follows from Lemma 3.5.1. For $\mathcal{M} = B_R(0)$, we use the bound $\mathcal{N}(B_R(0),\sigma_\delta/2) \leqslant \left(1+\frac{4R}{\sigma_\delta}\right)^d$ Vershynin, 2018, Corollary 4.2.13. □

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 3.5.2.

3.6 The effect of perturbing the data distribution on the score

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 3.5.

3.6.1 Perturbation under χ^2 error and truncation

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

Lemma 3.6.1 (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*

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

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

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

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

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

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

$$Q(\mathbb{R}^d) \geqslant 1-\varepsilon_{\mathrm{TV}}. \tag{3.32}$$

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 $\min\left\{\mathrm{d}QP_{0,y}, \mathrm{d}QP_{1,y}\right\}$. Let $\widehat{P}(d\widehat{x}_0, d\widehat{x}_1|y_0)$ denote this coupled distribution. Then as in Lemma 3.6.5,

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

We bound this by first bounding

$$\int_{\mathbb{R}^d} P_{0,1,y}(dy_1,dy_2)\widehat{P}(\widehat{x}_0\neq\widehat{x}_1) \leqslant \int_{\mathbb{R}^d} P_{0,y}(dy)\max\left\{1-\mathrm{d}QP_{0,y}, 1-\mathrm{d}QP_{1,y}\right\} \leqslant 2\varepsilon_{\mathrm{TV}}, \tag{3.33}$$

which follows from the two inequalities (using (3.32))

$$\int_{\mathbb{R}^d} P_{0,y}(dy)\left(1 - \mathrm{d}QP_{0,y}\right) = 1 - Q(\mathbb{R}^d) \leqslant \varepsilon_{\mathrm{TV}}$$

$$\begin{aligned}\int_{\mathbb{R}^d} P_{0,y}(dy)\left(1 - \mathrm{d}QP_{1,y}\right) &\leqslant \int_{\mathbb{R}^d} P_{1,y}(dy)\left(1 - \mathrm{d}QP_{1,y}\right) + \mathrm{TV}(P_{0,y}, P_{1,y}) \\ &\leqslant (1 - Q(\mathbb{R}^d)) + \varepsilon_{\mathrm{TV}} \leqslant 2\varepsilon_{\mathrm{TV}}.\end{aligned}$$

From (3.33), 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

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

Now for the second term (II),

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

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

$$\begin{aligned}\int_\Omega f(x)P(dx) &\leqslant \int_\Omega f(x)Q(dx) + \int_\Omega (\mathrm{d}PQ - 1)\, f(x)Q(dx) \\ &\leqslant \int_\Omega f(x)Q(dx) + \sqrt{\chi^2(P||Q)\int_\Omega f(x)^2 Q(dx)}\end{aligned}$$

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

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

$$\begin{aligned}&\leqslant \int_{\mathbb{R}^n} P_{1,y}(dy_0)P_{0,1,y}(y_0 \neq y_1|y_0)\,\|r_1(y_0)\|^2 \\ &\quad + \sqrt{\chi^2(P_{0,y}||P_{1,y})\int P_{1,y}(dy_0)P_{0,1,y}(y_0 \neq y_1|y_0)\,\|r_1(y_0)\|^4}.\end{aligned}$$

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

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

so

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

where we used the data processing inequality.

For $\varphi = \varphi_{\sigma^2}$, we obtain by Lemma 3.6.6 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 3.6.2 (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 $\chi^2(\widetilde{P}^{(1)}||\widetilde{P}^{(0)}) \leqslant \varepsilon_\chi^2 \leqslant 1$.*

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

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

2. *Let $\widetilde{p}_t^{(i)}$ be the density resulting from running* (3.1) *starting from $\widetilde{P}^{(i)}$, and let σ_t be as in* (3.3)*. 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 (3.28) 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

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

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 3.6.1 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 $\widetilde{P}^{(i)} \leftarrow M_{m_t \sharp} \widetilde{P}^{(i)}$ (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 3.6.2.

Lemma 3.6.3. *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*

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

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

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

Proof. We have

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

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

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

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

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

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 3.2.3.

Lemma 3.6.4. *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 $\alpha_t\sigma_t \leqslant \frac{1}{2L}$, we have*

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

Here $\alpha_t = 1/m_t$ and σ_t are defined in (3.3)*. In particular, $TV(\widetilde{p}_\delta, \widetilde{p}_0) \leqslant \varepsilon_{\mathrm{TV}}$ 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,

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

For the second term,

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

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

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

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:

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

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

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

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

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

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

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

which is implied by

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

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

3.6.2 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 3.6.5. *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*

- $\mathbb{E}[\|y_0 - y_1\|^2] \leqslant \varepsilon_W^2$.

Define the tail error by

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

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

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

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

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

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

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

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 3.6.2, 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 $\min\{\mathrm{d}QP_{0,y}, \mathrm{d}QP_{1,y}\}$. Let $\widehat{P}(d\widehat{x}_0, d\widehat{x}_1|y_0)$ denote this coupled

distribution. Then

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

as in Lemma 3.6.2. Now

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

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

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

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

For the second upper bound, we note that

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

3.6.3 Gaussian tail calculation

We use the following Gaussian tail calculation in the proof of Lemma 3.6.2.

Lemma 3.6.6. *Let μ be the standard Gaussian measure on $N(0, I_d)$. Then*

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

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

Proof. By the χ^2 tail bound in Laurent and Massart, 2000, for $t \geqslant 0$,

$$
\mu(\|X\|^2 \geqslant 2d + 3t) \leqslant \mathbb{P}(\|X\|^2 \geqslant d + 2\sqrt{dt} + 2t) \leqslant e^{-t}, \tag{3.34}
$$

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,

$$\sup_{\mu(A)\leqslant\varepsilon}\int_A \|x\|^2\,\mu(dx) \leqslant \sup_{P_Y(A)\leqslant\varepsilon}\int_A yP_Y(dy) = \int_{2d+3\ln\left(\frac{1}{\varepsilon}\right)}^{\infty} ydF(y)$$

$$= \varepsilon\left(2d+3\ln\left(\frac{1}{\varepsilon}\right)\right) + \int_{2d+3\ln\left(\frac{1}{\varepsilon}\right)}^{\infty} e^{-\frac{y-2d}{3}}dy = \varepsilon\left(2d+3\ln\left(\frac{1}{\varepsilon}\right)\right) + 3\varepsilon$$

and

$$\sup_{\mu(A)\leqslant\varepsilon}\int_A \|x\|^4\,\mu(dx) \leqslant \sup_{P_Y(A)\leqslant\varepsilon}\int_A y^2P_Y(dy) = \int_{2d+3\ln\left(\frac{1}{\varepsilon}\right)}^{\infty} y^2dF(y)$$

$$= \varepsilon\left(2d+3\ln\left(\frac{1}{\varepsilon}\right)\right)^2 + \int_{2d+3\ln\left(\frac{1}{\varepsilon}\right)}^{\infty} 2ye^{-\frac{y-2d}{3}}dy$$

$$= \varepsilon\left(2d+3\ln\left(\frac{1}{\varepsilon}\right)\right)^2 - \left[3ye^{-\frac{y-2d}{3}}\right]\Big|_{2d+3\ln\left(\frac{1}{\varepsilon}\right)}^{\infty} + \int_{2d+3\ln\left(\frac{1}{\varepsilon}\right)}^{\infty} 3e^{-\frac{y-2d}{3}}\,dy$$

$$= \varepsilon\left(2d+3\ln\left(\frac{1}{\varepsilon}\right)\right)^2 + 3\varepsilon\left(2d+3\ln\left(\frac{1}{\varepsilon}\right)\right) + 9\varepsilon. \qquad \square$$

3.7 Guarantees under L^2-accurate score estimate

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

Assumption 8 (Tail bound). *$R : [0,1] \to [0,\infty)$ is a function such that $p_{\text{data}}(B_{R(\varepsilon)}(0)) \geqslant 1-\varepsilon$.*

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 4) as a special case.

3.7.1 TV error guarantees

We follow the framework of Lee et al., 2022 to convert guarantees under L^∞-accurate score estimate, to guarantees under L^2-accurate score estimate.

Theorem 3.7.1 (Lee et al., 2022, 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 $0 \leqslant k \leqslant n-1$, then $Z_n = \overline{Z}_n$.*
2. $\chi^2(\overline{q}_n || p_n) \leqslant D_n^2$.
3. $\mathbb{P}(X_n \in B_n) \leqslant \delta_n$.

Then the following hold.

$$\mathrm{TV}(q_n, \overline{q}_n) \leqslant \sum_{k=0}^{n-1} (D_k^2 + 1)^{1/2} \delta_k^{1/2} \qquad \mathrm{TV}(p_n, q_n) \leqslant D_n + \sum_{k=0}^{n-1} (D_k^2 + 1)^{1/2} \delta_k^{1/2} \tag{3.35}$$

Theorem 3.7.2 (DDPM with L^2-accurate score estimate)**.** *Let $0 < \varepsilon_\chi, \varepsilon_{\mathrm{TV}}, \delta < \frac{1}{2}$. Suppose that Assumption 8 for a sufficiently small value of c that R_0 is such that $R\left(\frac{c\varepsilon_{\mathrm{TV}}^3 \delta^6 \varepsilon_\chi^{12}}{R_0^{19} d^5}\right) \leqslant R_0$, and $R_0^2 \geqslant d$. Suppose one of the following cases holds.*

1. *Let $p_{\mathrm{data}}, s(\cdot, t)$ be such that Assumption 3 holds, with $R_0^2 \geqslant d$. 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 (3.5) 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 $p_{\mathrm{data}}, s(\cdot, t)$ be such that Assumptions 3 and 5 hold, with $C \geqslant R_0^2$. 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 (3.5) 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 $\chi^2(\overline{q}_{t_N}||p_{t_N}) \leqslant \varepsilon_\chi^2$. In particular, taking $\varepsilon_\chi = \varepsilon_{\mathrm{TV}}$, we have $\mathrm{TV}(q_T, p_{\mathrm{data}}) \leqslant 2\varepsilon_{\mathrm{TV}}$.

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 3.5.2 for a ε_K to be chosen, to obtain a distribution $\widetilde{P}_0$ on $B_{R_0}(0)$, where $R_0 \geqslant R(\varepsilon_K/8)$. 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 $\mathrm{TV}(p_{\mathrm{data}}, \widetilde{P}_0) \leqslant \sqrt{\varepsilon_K} = o(\varepsilon_{\mathrm{TV}})$ (and the same holds for their evolutions under (3.1)), it suffices to consider convergence to $\widetilde{p}_\delta$.

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

$$B_t := \{\|\nabla \ln \widetilde{p}_t(x) - s(x,t)\| > \varepsilon_{\infty,t}\} \tag{3.36}$$

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

Given $t \geqslant 0$, 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

$$d\overline{z}_t = g(T-t)^2 \left(\frac{1}{2} z_t + b(z_-, T - t_-)\right) dt + g(T-t)\, dw_t \tag{3.37}$$

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

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 (3.13).

We can couple this process with the exponential integrator (3.5) using s so that as long as $x_{t_m} \notin B_{T-t_m}$, the processes agree, thus satisfying condition 1 of Theorem 3.7.1.

Then by Lemma 3.6.3, substituting the bound of Lemma 3.6.2 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 3.4.11 or 3.4.12, when $\int_0^{t_n} \varepsilon_t^2\, dt = O(1)$,

$$\begin{aligned}\chi^2(\overline{q}_{t_k}||p_{t_k}) &= e^{\varepsilon}\chi^2(q_0||p_0) + \varepsilon + e^{\varepsilon}\int_0^{t_n} \varepsilon_{\infty,T-t}^2\, dt \qquad (3.38)\\ &\leqslant 2\chi^2(q_0||p_0) + O(1),\end{aligned}$$

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 (3.38) to be bounded by $\frac{\varepsilon_\chi^2}{2}, \frac{\varepsilon_\chi^2}{4}, \frac{\varepsilon_\chi^2}{4}$; this is implied by

$$T = \ln\left(\frac{16R^2}{\varepsilon_\chi^2}\right) \text{ by Lemma 3.4.14}$$

$$\int_0^{t_n} \varepsilon_{\infty,T-t}^2\, dt = O(\varepsilon_\chi^2). \qquad (3.39)$$

By Theorem 3.7.1,

$$\begin{aligned}\mathrm{TV}(q_{t_n}, \overline{q}_{t_n}) &\leqslant \sum_{k=0}^{n-1}(1 + \chi^2(\overline{q}_{t_k}||p_{t_k}))^{1/2} P(B_{t_k})^{1/2}\\ &\leqslant \sum_{k=0}^{n-1}\left(2\chi^2(q_0||p_0)^{1/2} + O(1)\right)\delta_t^{1/2} \qquad (3.40)\\ &= O\left(\sum_{k=0}^{n-1}\frac{\varepsilon_{t_k}}{\varepsilon_{\infty,t_k}} + \sqrt{\varepsilon_K}\left(1 + \frac{\sqrt{d + \ln(1/\varepsilon_K)}}{\varepsilon_{\infty,t_k}\sigma_{T-t_k}}\right)\right). \qquad (3.41)\end{aligned}$$

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

$$\sum_{k=0}^{n-1}\frac{\varepsilon_t}{\varepsilon_{\infty,t}} = O(\varepsilon_{\mathrm{TV}}) \qquad (3.42)$$

$$\varepsilon_K = O\left(\frac{\min_k \varepsilon_{t_k}^2 \sigma_{T-t_k}^2}{d + \ln(1/\varepsilon_K)}\right). \qquad (3.43)$$

We bound (3.43) 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

$$\varepsilon_K = O\left(\frac{\varepsilon_\sigma^2}{d + \ln(1/\varepsilon_K)}\right). \tag{3.44}$$

We will return to this after deriving a condition on ε_σ. It remains to bound (3.39) and (3.42). 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 $t_{n^{\text{coarse}}-1} \leqslant T - 1 \leqslant t'_1$. 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 (3.39) and the sum (3.42) into the parts involving the coarse and fine timepoints. For (3.39), it suffices to have

$$\text{(3.39), coarse:} \quad \int_0^{t'_0} \varepsilon_{\infty,T-t}^2 \, dt \leqslant T \max_{0 \leqslant k \leqslant n^{\text{coarse}}} \varepsilon_{\infty,T-t_k}^2 = O(\varepsilon_\chi^2)$$

so it suffices to take $\varepsilon_{\infty,T-t_k}^2 \asymp \frac{\varepsilon_\chi^2}{T}$. 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)

$$\text{(3.39), fine:} \quad \int_{t'_0}^{t'_{n^{\text{fine}}}} \varepsilon_{\infty,T-t}^2 \, dt = \sum_{k=0}^{n^{\text{fine}}-1} h'_k \varepsilon_{\infty,T-t'_k}^2 = O(\varepsilon_\chi^2)$$

$$\Longleftarrow \sum_{k=0}^{n^{\text{fine}}-1} \frac{\varepsilon_\chi^2 u_k^\alpha}{B} \varepsilon_{\infty,u_k}^2 = O(\varepsilon_\chi^2)$$

$$\Longleftarrow \sum_{k=0}^{n^{\text{fine}}-1} \frac{u_k^\alpha \varepsilon_{\infty,u_k}^2}{B} = O(1). \tag{3.45}$$

For (3.42), it suffices to have

$$\text{(3.42), coarse:} \quad \sum_{k=0}^{n^{\text{coarse}}-1} \frac{\varepsilon_{T-t_k}}{\varepsilon_{\infty,T-t_k}} \asymp n^{\text{coarse}} \frac{\varepsilon_\sigma}{\varepsilon_\chi/\sqrt{T}} = O(\varepsilon_{\text{TV}})$$

$$\Longleftarrow \varepsilon_\sigma = O\left(\frac{\varepsilon_{\text{TV}} \varepsilon_\chi}{n^{\text{coarse}}\sqrt{T}}\right) \tag{3.46}$$

and

$$\text{(3.42), fine:} \quad \sum_{k=0}^{n^{\text{fine}}-1} \frac{\varepsilon_{u_k}}{\varepsilon_{\infty,u_k}} \asymp \sum_{k=0}^{n^{\text{fine}}-1} \frac{\varepsilon_\sigma}{u_k \varepsilon_{\infty,u_k}} = O(\varepsilon_{\text{TV}}). \tag{3.47}$$

Note that in light of the required step sizes, we can take $n^{\text{coarse}} \asymp \frac{T^2 B}{\varepsilon_\chi^2}$. Considering the equality case of Hölder's inequality on $(3.45)^{1/3}(3.47)^{2/3}$ suggests that we take

$$\varepsilon_\sigma \asymp \frac{\varepsilon_{\text{TV}} B^{1/2}}{\left(\sum_{k=0}^{n^{\text{fine}}-1} {u_k}^{\frac{\alpha-2}{3}}\right)^{3/2}} \tag{3.48}$$

$$\varepsilon_{\infty,u_k} \asymp \frac{B^{1/2}}{{u_k}^{\frac{\alpha+1}{3}} \left(\sum_{k=0}^{n^{\text{fine}}-1} {u_k}^{\frac{\alpha-2}{3}}\right)^{1/2}} \tag{3.49}$$

Note that the number of steps needed in the fine part is $O\left(\frac{B}{\varepsilon_\chi^2 \delta^2}\right)$ in the first case and $O\left(\frac{B}{\varepsilon_\chi^2}\right) \ln\left(\frac{1}{\delta}\right)$ in the second case. We can check that (3.48) and (3.49) make (3.45) and (3.47) 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 $h_k' \asymp \frac{\varepsilon_\chi^2}{B(T-t_k')^3}$, it takes $n^{\text{fine}} = \Theta\left(\frac{B}{\varepsilon_\chi^2 \delta^2}\right)$ steps to reach $T - t = \delta$.

$$u_k = T - t_k' = \left(1 + \Theta\left(\frac{k\varepsilon_\chi^2}{B}\right)\right)^{-\frac{1}{2}}$$

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

Then we obtain

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

In case 1, our requirement is

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

but note that the first bound is more stringent. Now, returning to (3.44), 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 3.7.3 (TV error for DDPM with L^2-accurate score estimate and smoothness)**.** *Let $0 < \varepsilon_{\mathrm{TV}} < \frac{1}{2}$. Suppose that Assumption 6 and 8 for a sufficiently small value of c that R_0 is such that $R\left(\frac{c\varepsilon_{\mathrm{TV}}^{15}}{R_0^{31}d^5L^{12}}\right) \leqslant R_0$, and $R_0^2 \geqslant \max\left\{d, \mathbb{E}_{P_{data}}\left[\|X\|^2\right]\right\}$, and one of the following cases holds.*

1. *Let $p_{\mathrm{data}}, s(\cdot,t)$ be such that Assumption 3 holds. Suppose that*

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

where $B = R_0^4 d\ln\left(\frac{TR_0^2L^2}{\varepsilon_{\mathrm{TV}}^2}\right)\ln\left(\frac{R_0^3dL^2}{\varepsilon_{\mathrm{TV}}^3\varepsilon_\chi}\right)$, and we run (3.5) starting from p_{prior} for time $T = \ln\left(\frac{16R_0^2}{\varepsilon_{\mathrm{TV}}^2}\right)$, $N = O\left(\frac{B\left(T+\left(\frac{R_0L}{\varepsilon_{\mathrm{TV}}}\right)^4\right)}{\varepsilon_{\mathrm{TV}}^2}\right)$ 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 $p_{\mathrm{data}}, s(\cdot,t)$ be such that Assumptions 3 and 5 hold, with $C \geqslant R_0^2$. Suppose*

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

where $B = C^2 d\ln\left(\frac{TR_0^2L^2}{\varepsilon_{\mathrm{TV}}^2}\right)\ln\left(\frac{R_0^3dL^2}{\varepsilon_{\mathrm{TV}}^4}\right)$, and we run (3.5) starting from p_{prior} for time $T = \ln\left(\frac{16R_0^2}{\varepsilon_{\mathrm{TV}}^2}\right)$, $N = O\left(\frac{B\left(T+\ln\left(\frac{R_0L}{\varepsilon_{\mathrm{TV}}}\right)\right)}{\varepsilon_{\mathrm{TV}}^2}\right)$ 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 ε_{TV}-far in TV distance from the data distribution P_{data}.

Proof. With the result of Theorem 3.7.2, we see that $\text{TV}(q_{t_N}, p_{t_N}) \leqslant 2\varepsilon_{\text{TV}}$. Now by Lemma 3.6.4, if we further assume

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

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

Proof of Theorem 3.2.3. If p_{data} is subexponential with a fixed constant, note that Assumption 8 holds with $R(\varepsilon) = O\left(\ln\left(\frac{1}{\varepsilon}\right)\right)$ and hence R_0 is logarithmic in all parameters. □

3.7.2 Wasserstein error guarantees

Proof of Theorem 3.2.1. Suppose $T - t_N = \delta$. Note by (3.3) 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 $W_2(\widetilde{p}_0, M_{m_\delta^{-1}\#}\widetilde{p}_\delta) \leqslant m_\delta^{-1}\sigma_\delta\sqrt{d} \leqslant \sqrt{e^\delta - 1} \leqslant \sqrt{2\delta d}$ (for $\delta \leqslant 1$). Choosing $\delta = \frac{\varepsilon_{\text{W}}^2}{2d}$, we see by Theorem 3.7.2 it suffices to take

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

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

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

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

Proof of Theorem 3.2.2. 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,

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

We choose $\delta = \frac{\varepsilon_{\mathrm{W}}^2}{8d}$ so the first term is $\leqslant \frac{\varepsilon_{\mathrm{W}}}{2}$. 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

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

where the bound on the second term uses Lemma 3.6.6. Using $R^2 \geqslant d$, 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 3.7.2, 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 $\varepsilon_\sigma = \widetilde{o}\left(\frac{\varepsilon_{\mathrm{W}}^8}{C^2 R^8 d}\right)$. □

3.8 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 3.4.13, note that we can express (3.28) and (3.29) as

$$\nabla \ln(\mu * \varphi_{\sigma^2}(y)) = -\frac{1}{\sigma^2}\mathbb{E}[Y - X|Y = y] \tag{3.50}$$

$$\nabla^2 \ln(\mu * \varphi_{\sigma^2}(y)) = \frac{1}{\sigma^4}\operatorname{Cov}[Y - X|Y = y] - \frac{1}{\sigma^2} I_d \tag{3.51}$$

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 (3.51) 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 3.4.13 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 3.8.1. *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*

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

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

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

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 $\leqslant 5^d$ Vershynin, 2018, Cor. 4.2.13. By a union bound,

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

when we take $\lambda = \|X\|_{\psi_2} \sqrt{\ln\left(\frac{2\cdot 5^d}{\varepsilon}\right)}$. By Vershynin, 2018, Lemma 4.4.1, the operator norm can be bounded by the norm on an ε-net,

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

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

From this we obtain the desired high-probability bound.

Lemma 3.8.2. *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* (3.1) *at time t, we have*

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

Note that there is no dependence on the radius.

Proof. Apply (3.51) 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 $\|Y - X\|_{\psi_2} \leqslant C_2 \sigma$ for some universal constant C_2, the result follows from Lemma 3.8.1. □

4. Convergence of flow-based generative mode

4.1 Background and Overview

Flow-based generative models enjoy certain advantages in computing the data generation and the likelihood, and have recently shown competitive empirical performance. Compared to the accumulating theoretical studies on related score-based diffusion models, analysis of flow-based models, which are deterministic in both forward (data-to-noise) and reverse (noise-to-data) directions, remain sparse. In this chapter, we provide a theoretical guarantee of generating data distribution by a "progressive" flow model, mainly following the JKO flow model in Xu et al., 2023b but similar models have been proposed in, e.g., Alvarez-Melis et al., 2021; Mokrov et al., 2021; Vidal et al., 2023. We prove the exponential convergence rate of such flow models in both (data-to-noise and noise-to-data) directions. We start by briefly discussing several types of flow-based models, particularly the progressive one, and then introduce the main results of this chapter. A more complete literature survey can be found in Section 4.1.1.

Normalizing flow. Normalizing flow is a class of deep generative models for efficient sampling and density estimation. Compared to diffusion models, Continuous Normalizing Flow (CNF) models (Kobyzev et al., 2020) appear earlier in the generative model literature. Largely speaking, CNF (flow-based models) fall into two categories: discrete-time and continuous-time. The discrete-time CNF models adopt the structure of a Residual Network (ResNet) (He et al., 2016) and typically consist of a sequence of mappings:

$$x_l = x_{l-1} + f_l(x_{l-1}), \quad l = 1, \cdots, L \tag{4.1}$$

where f_l is the neural network mapping parameterized by the l-th "Residual Block", and x_l is the output of the l-th block. Continuous-time CNFs are implemented under the neural ODE framework (R. T. Chen et al., 2018), where the neural network features $x(t)$ is computed by integrating an ODE

$$\dot{x}(t) = v(x(t), t), \quad t \in [0, T], \tag{4.2}$$

and $v_t(x) = v(x,t)$ is parametrized by a neural ODE network. The discrete-time CNF (4.1) can be viewed as computing the numerical integration of the neural ODE (4.2) on a sequence of time stamps via the forward Euler scheme.

In both categories, a CNF model computes a deterministic transport from the data distribution towards a target distribution q typically normal, $q = \mathcal{N}(0, I_d)$, per the name "normalizing". The forward time flow is illustrated in Figure 4.1. Taking the continuous-time formulation (4.2), let P be the data distribution with density p, $x(0) \sim p$, and denote by $p_t(x) = p(x,t)$ the probability density of $x(t)$. Then p_t solve the continuity equation (CE)

$$\partial_t p_t + \nabla \cdot (p_t v_t) = 0, \tag{4.3}$$

from $p_0 = p$. If the algorithm can find a v_t such that p_T at some time T is close to q, then one would expect the reverse-time flow from $t = T$ to $t = 0$ to transport from q to a distribution close to p. Note that in the continuous-time flow, invertibility is presumed since the neural ODE can be integrated in two directions of time alike. For discrete-time flow (4.1), invertibility needs to be ensured either by special designs of the neural network layer type (Dinh et al., 2014, 2016; Kingma & Dhariwal, 2018), or by regularization techniques such as spectral normalization (Behrmann et al., 2019) or transport cost regularization (Onken et al., 2021; Xu et al., 2022).

A notable advantage of the flow model is the computation of the likelihood. For discrete-time flow (4.1), this involves the computation of the log-determinant of the Jacobian of f_l. For continuous-time flow (4.2), this is by the relation

$$\log p_t(x(t)) - \log p_s(x(s)) = -\int_s^t \nabla \cdot v(x(\tau), \tau) d\tau,$$

which involves the time-integration of the trace of the Jacobian of v_t (Grathwohl et al., 2018). While these computations may encounter challenges in high dimensions, the ability to evaluate the (log) likelihood is fundamentally useful; in particular, it allows for evaluating the maximum likelihood training objective on finite samples. This property is also adopted in the deterministic reverse process in diffusion models (Song, Sohl-Dickstein, et al., 2021),

called the "probability flow ODE" (see more in Section 4.1.1), so the likelihood can be evaluated once a forward diffusion model has been trained.

Progressive flow models. Another line of works, developed around the same time as diffusion models, explored the variational form of normalizing flow as a Wasserstein gradient flow and proposed *progressive* training of the flow model. The progressive training of ResNet, i.e., training block-wise by a per-block variational loss, was proposed by Johnson and Zhang, 2019 at an earlier time under the GAN framework. Later on, the Jordan-Kinderleherer-Otto (JKO) scheme, as a time-discretized Wasserstein gradient flow (see more in Section 4.2.3), was explored in several flow-based generative models: Alvarez-Melis et al., 2021; Mokrov et al., 2021 implemented the JKO scheme using input convex neural networks (Amos et al., 2017); Fan et al., 2022 proposed a forward progressive flow from noise to data, showing empirical success in generating high-dimensional real datasets; Xu et al., 2023b developed the JKO flow model under the invertible continuous-time CNF framework, achieving competitive generating performance on high-dimensional real datasets at a significantly reduced computational and memory cost from previous CNF models; an independent concurrent work Vidal et al., 2023 proposed a block-wise JKO flow model utilizing the framework of Onken et al., 2021. Many other flow models related to diffusion and Optimal Transport (OT) exist in the literature; see more in Section 4.1.1. Our theoretical analysis will focus on the progressive flow models, and we primarily follow the invertible flow framework in Xu et al., 2023b.

To be more specific, a progressive flow model represents the flow on $[0, T]$ as the composition of N sub-flow models where each one computes the flow on a sub-interval $[t_{n-1}, t_n]$, $n = 1, \ldots, N$. The training is "progressive", meaning that at one time, only one sub-model is trained, and the training of the n-th sub-model is conducted once the previous $n-1$ sub-models are trained and fixed. The sub-flow model on $[t_{n-1}, t_n]$ can take different forms, e.g., a ResNet block or a continuous-time neural ODE, and the N sub-intervals always provide a time-discretization of the flow. In this context, we call the sub-model on

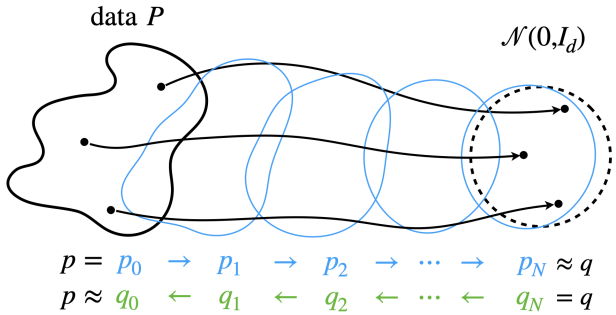


FIGURE 4.1: The forward and reverse processes (4.19) consist of the sequence of transported densities at discrete time stamps.

the n-th sub-interval a "Residual Block".

Overview of the main result. An abundance of theoretical works has provided the generation guarantee of score-based diffusion models (Benton, De Bortoli, et al., 2023; H. Chen et al., 2023; S. Chen et al., 2022; De Bortoli, 2022; Lee et al., 2022, 2023; Pedrotti et al., 2023). In comparison, the theoretical study of flow-based generative models is much less developed. Most recent works on the topic focused on the generation guarantee of the ODE reverse process (deterministic sampler) once a score-based model is trained from the *forward* SDE diffusion process (S. Chen, Chewi, et al., 2023; S. Chen, Daras, & Dimakis, 2023; Li et al., 2023). For generative models which are flow-based in the forward process, generation guarantee for flow-matching models under continuous-time formulation was shown in Albergo and Vanden-Eijnden, 2023 under $\mathcal{W}_2$, and in Albergo et al., 2023 under the Kullback–Leibler (KL) divergence by incorporating additional SDE diffusion. This chapter focuses on obtaining the theoretical guarantee of the JKO flow model (Xu et al., 2023b), which is progressively trained over the Residual Blocks (steps) and generates a discrete-time flow in both forward and reverse directions. The mathematical formulation of the JKO flow is summarized in Section 4.3, where we introduce needed theoretical assumptions on the learning procedure.

Our analysis is based on first proving the convergence of the forward process (the JKO scheme by flow network), which can be viewed as an approximate proximal Gradient Decent (GD) in the Wasserstein-2 space to minimize $G(\rho)$, a functional on the space of

probability distributions. While the convergence analyses of Wasserstein GD and proximal GD have appeared previously in literature Kent et al., 2021; Salim et al., 2020, our setup differs in several ways, primarily in that we consider the JKO scheme, which is a "fully-backward" discrete-time GD. For the N step discrete-time proximal GD, which produces a sequence of transported distributions p_n, we prove the convergence of both $\mathcal{W}_2(p_n, q)$ and the objective gap $G(p_n) - G(q)$ at an exponential rate, where q is the global minimum of G (Theorem 8). The convergence applies to a general class of (strongly) convex G that includes the KL divergence $\mathrm{KL}(p||q)$ as a special case. This result echos the classical proximal GD convergence in vector space where one expects an exponential convergence rate for strongly convex minimizing objectives. While exponential convergence is a natural result from the point of view of gradient flow, this convergence result of JKO-type $\mathcal{W}_2$-proximal GD did not appear in previous literature to the authors' best knowledge and can be of independent interest.

After obtaining a small $G(p_n) = \mathrm{KL}(p_n||q)$ from the convergence of the forward process, we directly obtain the KL guarantee of the generated density from the data density by the invertibility of the flow and the data processing inequality, and this implies the total variation (TV) guarantee (Corollary 10). The requirement for data distribution is having a finite second moment and a density (with respect to the Lebesgue measure). The TV and KL guarantees are of $O(\varepsilon)$ and $O(\varepsilon^2)$, respectively, where ε is the bound for the magnitude of the Wasserstein (sub-)gradient of the loss function (hence error in the first order condition) in each of the N JKO steps (Assumption 1), and the process achieves the error bound in $N \lesssim \log(1/\varepsilon)$ many steps (each step is a Residual Block).

To handle the situation when the data distribution only has a finite second moment but no density, we apply a short-time initial diffusion and start the forward process from the smoothed density. This short-time diffusion was adopted in practice and prior theoretical works. We then obtain KL and TV guarantee to generate the smoothed density, which can be made arbitrarily close to the data distribution in $\mathcal{W}_2$ when the initial diffusion time duration tends to zero (Corollary 11). The above results are obtained when the reverse

process is computed exactly with no inversion error. Our analysis can also extend to the case of small inversion error by proving a $\mathcal{W}_2$-guarantee between the generated density from the exact reverse process and that from the actual computed one. Theoretically, the $\mathcal{W}_2$-error can be made $O(\varepsilon)$ or smaller assuming that the inversion error can be made $O(\varepsilon^\alpha)$ for some exponent α (Corollary 13).

4.1.1 Related works

Score-based diffusion models

In score-based diffusion models, the algorithm first simulates a forward process, which is a (time-discretized) SDE, from which a score function parameterized as a neural network is trained. The reverse (SDE or ODE) process is simulated using the learned score model to generate data samples.

SDE in diffusion models. As a typical example, in the *variance preserving* Denoising Diffusion Probabilistic Modeling (DDPM) process (Ho et al., 2020; Song, Sohl-Dickstein, et al., 2021), the forward process produces a sequence of X_n,

$$X_n = \sqrt{1-\beta_n} X_{n-1} + \sqrt{\beta_n} Z_{n-1}, \quad n = 1, \cdots, N, \quad Z_n \sim \mathcal{N}(0, I_d) \text{ i.i.d.} \tag{4.4}$$

where $X_0 \sim P$ is drawn from data distribution, With large N, the continuum limit of the discrete dynamic (4.4) is a continuous-time SDE

$$dX_t = -\frac{1}{2}\beta(t) X_t dt + \sqrt{\beta(t)} dW_t, \quad t \in [0, T], \tag{4.5}$$

where $\beta(t) > 0$ is a function and W_t is a standard Wiener process (Brownian motion). Since $\beta(t)$ in (4.5) corresponds to a time reparametrization of t, after changing the time ($t \mapsto \int_0^t \beta(s)/2 ds$), (4.5) becomes the following SDE

$$dX_t = -X_t dt + \sqrt{2} dW_t, \tag{4.6}$$

which is the Ornstein-Uhlenbeck (OU) process in $\mathbb{R}^d$. We consider the time-parametrization in (4.6) for exhibition simplicity. More generally, one can consider a diffusion process

$$dX_t = -\nabla V(X_t)\, dt + \sqrt{2}\, dW_t, \quad X_0 \sim P, \tag{4.7}$$

and the OU process is a special case with $V(x) = \|x\|^2/2$.

We denote by $\rho_t = \mathcal{L}_t(P)$ the marginal distribution of X_t for $t > 0$. The time evolution of ρ_t is described by the Fokker–Planck Equation (FPE) written as

$$\partial_t \rho_t = \nabla \cdot (\rho_t \nabla V + \nabla \rho_t). \tag{4.8}$$

Forward and reverse processes. When simulating the forward process, the diffusion models train a neural network to learn the score function $s_t(x) := \nabla \log \rho_t$ by score matching (Hyvärinen & Dayan, 2005; Vincent, 2011). The training objective can be expressed as the mean-squared error defined as $\int_0^T \int \|\hat{s}_t(x) - s_t(x)\|^2 \rho_t(x) dx dt$, which facilitates training and is scalable to high dimension data such as images (in the original pixel space).

Once the neural-network score function $\hat{s}_t$ is learned, the algorithm simulates a reverse-time SDE $\tilde{X}_t$ (with time discretization in practice) (Song, Sohl-Dickstein, et al., 2021), such that from $\tilde{X}_T \sim \mathcal{N}(0, I)$ the distribution of $\tilde{X}_0$ is expected to the close to the data distribution P. It has also been proposed in Song, Sohl-Dickstein, et al., 2021 to compute the reverse process by integrating the following ODE (corresponding to (4.6))

$$\dot{\tilde{x}}(t) = -\nabla V(\tilde{x}(t)) - s_t(\tilde{x}(t)) \tag{4.9}$$

reverse in time, and (4.9) was called the "probability flow ODE". The validity of this ODE reverse process can be justified by the observation that the CE (4.3) and FPE (4.8) are the same when setting $v_t(x) = -(\nabla V(x) + s_t(x))$. This equivalence between density evolutions by SDE and ODE has been known in the literature of diffusion processes and solving FPE, dating back to the 90s (Degond & Mas-Gallic, 1989; Degond & Mustieles, 1990).

Flow models related to diffusion and OT

Flow-matching models. After diffusion models gained popularity, several flow-based models (in the reverse and forward directions) closely related to the diffusion model emerged. In particular, the Flow-Matching ODE model was proposed in Lipman et al., 2023 using the formulation conditional probability paths, where a neural ODE parameterized $\hat{v}(x, t)$ is trained to match a velocity field $v(x, t)$ whose corresponding CE (4.3) can evolve the density

p_t towards normality. The algorithm can adopt diffusion paths, where the CE will equal the density evolution equation (4.8) of an SDE forward process, as well as non-diffusion paths. A similar approach was developed under the "stochastic interpolant" framework in Albergo and Vanden-Eijnden, 2023, where the terminal distribution q can be arbitrary (not necessarily the normal distribution) and only accessible via finite samples. These models train a continuous-time CNF by minimizing a "matching" objective instead of the maximum likelihood objective as in Grathwohl et al., 2018, thus avoiding the computational challenges of the latter.

Optimal Transport flows. Apart from diffusion models and Wasserstein gradient flow, Wasserstein distance and OT have inspired another line of works on flow models where the Wasserstein distance, or a certain form of transport cost, is used to regularize the flow model and to compute the OT map between two distributions. Transport cost regularization of neural network models was suggested in several places: Ruthotto et al., 2020 provided a general framework for solving high-dimensional mean-field games (MFG) and control problems, Finlay et al., 2020 proposed a kinetic regularization aiming to stabilize neural ODE training, Onken et al., 2021 and Xu et al., 2022 developed the transport regularization in CNF and invertible ResNet, respectively, and H. Huang et al., 2023 applied to MFG and flow models. Other works developed flow models to compute the optimal coupling or the optimal transport between two distributions. For example, Rectified Flow Liu, 2022 proposed an iterative method to adjust the flow towards the optimal coupling. The method is closely related to the stochastic interpolant approach Albergo and Vanden-Eijnden, 2023 which, in principle, can solve the OT trajectory if the interpolant map can be optimized. A flow model to compute the dynamic OT between two high dimensional distributions from data samples was proposed in Xu et al., 2023a by refining the flow using the transport cost from a proper initialization. Despite the wealth of methodology developments and empirical results, the theoretical guarantees of these flow models are yet to be developed.

Theoretical guarantees of generative models

Approximation and estimation of GAN. On theoretical guarantees of generative models, earlier works focused on the approximation and estimation analysis under the GAN framework. The expressiveness of a deep neural network to approximate high dimensional distributions was established in a series of works, e.g., Lee et al., 2017; Lu and Lu, 2020; Perekrestenko et al., 2021; Yang et al., 2022, among others. The neural network architectures in these universal approximation results are typically feed-forward, like the generator network (G-net) proposed in the original GAN. The approximation and estimation of the discriminator network (D-net) in GAN were studied in Cheng and Cloninger, 2022, and the problem can be cast and analyzed as the learning of distribution divergences in high dimension Sreekumar and Goldfeld, 2022. Convergence analysis of GAN was studied in several places, e.g., J. Huang et al., 2022.

Guarantees of diffusion models. An earlier work Tzen and Raginsky, 2019 studied the expressiveness of a generative model using a latent diffusion process and proved guarantees for sampling and inference; however, the approach only involves a forward process and differs from the recent diffusion models. Motivated by the prevailing empirical success of score-based diffusion models, recent theoretical works centralized on the generation guarantee of such models using both SDE and ODE samplers, i.e., the reverse process.

For the SDE reverse process, the likelihood guarantee of the score-based diffusion model was first derived in Song, Durkan, et al., 2021 without time discretization. Taking into account the time discretization, which significantly influences the efficiency in practice, a series of theoretical studies have established polynomial convergence bounds for such models (Benton, De Bortoli, et al., 2023; H. Chen et al., 2023; S. Chen et al., 2022; De Bortoli, 2022; Lee et al., 2022, 2023; Pedrotti et al., 2023). In particular, Lee et al., 2022 were the first to attain polynomial convergence without succumbing to the curse of dimensionality, although this required a log-Sobolev Inequality on the data distribution. For a general data distribution, S. Chen et al., 2022 achieved polynomial error bounds in Total Variation (TV) distance

under the Lipschitz assumption, leveraging Girsanov's theorem. In parallel, Lee et al., 2023 derived similar polynomial convergence bounds, employing a technique for converting L^∞- accurate score estimates into L^2-accurate score estimation. Further advancements by H. Chen et al., 2023 established a more refined bound, reducing the requirement of smoothness of data distribution. Most recently, Pedrotti et al., 2023 improved the convergence rates under mild assumptions by introducing prediction-correction, and Benton, De Bortoli, et al., 2023 established the first convergence bounds for diffusion models, which are linear in the data dimension (up to logarithmic factors) without requiring any smoothness of the data distribution.

Guarantees of ODE flows. Within the studies of score-based diffusion models (note that the forward process is always SDE), theoretical findings for the ODE reverse process are relatively fewer. To the best of our knowledge, S. Chen, Daras, and Dimakis, 2023 established the first non-asymptotic polynomial convergence rate where the error bound involves an exponential factor in the flow time; S. Chen, Chewi, et al., 2023 provided the first polynomial-time convergence guarantees for the probability flow ODE implementation with a corrector step. Recently, Li et al., 2023 established bounds for both deterministic (ODE) and non-deterministic (SDE) samplers under certain additional assumptions on learning the score. The analysis is done by directly tracking the density ratio between the law of the diffusion process and that of the generated process in discrete time, leading to various non-asymptotic convergence rates.

Compared to score-based diffusion models, the guarantees of flow models (in both forward and reverse processes) are significantly less developed. We are aware of two recent works: The error bounds for the flow-matching model Albergo and Vanden-Eijnden, 2023 were proved in Benton, Deligiannidis, and Doucet, 2023 and applied to probability flow ODE in score-based diffusion models; for neural ODE models trained by likelihood maximization (the framework in Grathwohl et al., 2018), Marzouk et al., 2023 proved non-parametric statistical convergence rates to learn a distribution from data. Both works

used a continuous-time formulation, and the flow models therein are trained end-to-end. Compared to end-to-end training, progressive flow models can have advantages in training efficiency and accuracy, in addition to other advantages like smaller model complexity. For the analysis, the formulation of progressive flow models is variational and time-discretized in nature. Theoretical studies of time-discretized ODE flow models in both forward and reverse directions remain rudimentary.

Optimization in Wasserstein space

Continuing the classical literature in optimization and information geometry, several recent works established a convergence guarantee of first-order optimization in probability space in various contexts, leveraging the connection to the Wasserstein gradient flow. These analyses can potentially be leveraged under the theoretical framework of this paper to develop new (progressive) flow models as well as theoretical guarantees of generative models.

Optimization in probability distribution space. Convergence and rate analysis for first-order methods for vector-space optimization, primarily gradient descent and stochastic gradient descent - sometimes referred to as the Sample Average Approximation (SAA) approach - for convex and strongly convex problems have been established in the original works Nemirovski et al., 2009; Nemirovsky and Yudin, 1983, and extended in various contexts in subsequent papers. Optimization in the space of probability distributions (which forms a manifold) naturally arises in many learning problems and has become an important field of study in statistics and machine learning. In particular, the seminal work of Amari Amari, 2008, 2016 introduced information geometry emerging from studies of a manifold of probability distributions. It includes convex analysis and its duality as a special but important component; however, the line of work did not develop error analysis or convergence rates for algorithms on the probabilistic manifold. More recently, a Frank-Wolfe procedure in probability space was proposed in Kent et al., 2021 motivated by applications in nonparametric estimation and was shown to converge exponentially fast under general

mild assumptions on the objective functional.

Wasserstein proximal gradient descent. The landmark work Jordan et al., 1998 showed the solution to the Fokker Planck equation as the gradient flow of the KL divergence under the $\mathcal{W}_2$-distance. The proof in Jordan et al., 1998 employed a time discretization of the gradient flow now recognized as the JKO scheme. Making a connection between Langevin Monte Carlo and Wasserstein gradient flow, Bernton, 2018 proposed a proximal version of the Unadjusted Langevin Algorithm corresponding to a splitting scheme of the discrete Wasserstein GD and derived non-asymptotic convergence analysis. To analyze the convergence of discrete-time $\mathcal{W}_2$-gradient flow, Salim et al., 2020 introduced a Forward-Backward time discretization in the proximal Wasserstein GD and proved convergence guarantees akin to the GD algorithm in Euclidean spaces. We comment on the difference between Salim et al., 2020 and our scheme in more detail later in Remark 4.4.1.

The JKO scheme also inspired recent studies in variational inferences (VI). In the context of Gaussian VI, Lambert et al., 2022 proposed gradient flow of the KL divergence on the Bures-Wasserstein (BW) space, namely the space of Gaussian distributions on $\mathbb{R}^d$ endowed with the $\mathcal{W}_2$-distance. The algorithm enjoys the explicit solution of the JKO scheme in the BW space, and convergence of the continuous-time gradient flow was proved. In a follow-up work Diao et al., 2023, the forward-backward splitting was adopted in the proximal Wasserstein GD in the BW space, leading to convergence guarantees of the discrete-time GD to first-order stationary solutions. The closed-form solution of JKO operator only applies to the BW space, while the JKO flow network tries to learn a transport map to solve the JKO scheme in each step, leveraging the expressiveness of neural networks. Theoretically, we consider distributions with finite second moments in this work.

4.1.2 Notations

Throughout this chapter, we consider distributions over $\mathcal{X}$ and the domain $\mathcal{X} = \mathbb{R}^d$. We denote by $\mathcal{P}_2$, meaning $\mathcal{P}_2(\mathbb{R}^d)$, the space of probability distributions on $\mathbb{R}^d$ that has finite second moment. Specifically, for a distribution P, define $M_2(P) := \int_{\mathbb{R}^d} \|x\|^2 dP(x)$.

When P has a density (with respect to the Lebesgue measure dx), we also write $M_2(P)$ as $M_2(p)$. Then $\mathcal{P}_2 = \{P \text{ on } \mathbb{R}^d, s.t., M_2(P) < \infty\}$. We denote by $\mathcal{P}_2^r$ the distributions in $\mathcal{P}_2$ that have densities, namely $\mathcal{P}_2^r = \{P \in \mathcal{P}_2, P \ll dx\}$. We also say a density $p \in \mathcal{P}_2^r$ when $dP(x) = p(x)dx$ is in $\mathcal{P}_2^r$. In this paper, we consider distributions that have densities in most places. When there is no confusion, we use the density p to stand for both the density and the distribution $dP(x) = p(x)dx$, e.g., we say that a random variable $X \sim p$.

Given a (measurable) map $v : \mathbb{R}^d \to \mathbb{R}^d$ and P a distribution on $\mathbb{R}^d$, its L^2 norm is denoted as $\|v\|_P := (\int_{\mathbb{R}^d} \|v(x)\|^2 dP(x))^{1/2}$. When P has density p, we also denote it as $\|v\|_p$. For $P \in \mathcal{P}_2$, we denote by $L^2(P)$ (and also by $L^2(p)$ when P has density p) the L^2 space of vector fields, that is, $L^2(P) := \{v : \mathbb{R}^d \to \mathbb{R}^d, \|v\|_P < \infty\}$. For $u, v \in L^2(P)$, define $\langle u, v\rangle_P := \int_{\mathbb{R}^d} u(x)^T v(x) dP(x)$, which is also denoted as $\langle u, v\rangle_p$ when p is the density. The notation $\mathrm{I_d}$ stands for the identity map, which is always in $L^2(P)$ for $P \in \mathcal{P}_2$. For $T : \mathbb{R}^d \to \mathbb{R}^d$, the *pushforward* of a distribution P is denoted as $T_\# P$, such that $T_\# P(A) = P(T^{-1}(A))$ for any measurable set A. When P has density p and $T_\# P$ also has a density, we also denote by $T_\# p$ the density of $T_\# P$. For two maps $S, T : \mathbb{R}^d \to \mathbb{R}^d$, $S \circ T$ is the function composition.

4.2 Preliminaries

4.2.1 Wasserstein-2 distance and optimal transport

We first review the definitions of the Wasserstein-2 distance and optimal transport (OT) map, which are connected by the Brenier Theorem (see, e.g., Ambrosio et al., 2005, Section 6.2.3).

Given two distributions $\mu, \nu \in \mathcal{P}_2$, the Wasserstein-2 distance $\mathcal{W}_2(\mu, \nu)$ is defined as

$$\mathcal{W}_2^2(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, d\pi(x, y), \tag{4.10}$$

where $\Pi(\mu, \nu)$ denotes the family of all joint distributions with μ and ν as marginal distributions. When P and Q are in $\mathcal{P}_2^r$ and have densities p and q respectively, we also denote $\mathcal{W}_2(P, Q)$ as $\mathcal{W}_2(p, q)$. When at least one of μ and ν has density, we have the Brenier Theorem, which allows us to define the optimal transport (OT) map from μ to ν.

Theorem 1 (Brenier Theorem). *Let $\mu \in \mathcal{P}_2^r$ and $\nu \in \mathcal{P}_2$. Then*

(i) There exists a unique minimizer π of (4.10)*, which is characterized by a uniquely determined μ-a.e. map $T_\mu^\nu : \mathbb{R}^d \to \mathbb{R}^d$ such that $\pi = (\mathrm{I_d}, T_\mu^\nu)_\# \mu$, where $(\mathrm{I_d}, T_\mu^\nu)$ maps (x, y) to $(x, T_\mu^\nu(y))$. Moreover, there exists a convex function $\varphi : \mathbb{R}^d \to \mathbb{R}$ such that $T_\mu^\nu = \nabla \varphi$ μ-a.e.*

(ii) The minimum of (4.10) *equals that of the Monge problem, namely*

$$\mathcal{W}_2^2(\mu, \nu) = \inf_{T:\mathbb{R}^d \to \mathbb{R}^d,\, T_\# \mu = \nu} \int \|x - T(x)\|^2 d\mu(x).$$

(iii) If moreover $\nu \in \mathcal{P}_2^r$, then we also have the OT map T_ν^μ defined ν-a.e., and $T_\nu^\mu \circ T_\mu^\nu = \mathrm{I_d}$ μ-a.e., $T_\mu^\nu \circ T_\nu^\mu = \mathrm{I_d}$ ν-a.e.

In most places in our analysis, we will consider the *OT* between μ and ν both in $\mathcal{P}_2^r$, and we will frequently use the Brenier Theorem (iii) to obtain the pair of OT maps which are inverse of each other in the a.e. sense.

4.2.2 Differential and convexity of functionals on $\mathcal{P}_2$

Consider a proper lower semi-continuous functional $\phi : \mathcal{P}_2 \to (-\infty, \infty]$ and we denote the domain to be $\mathrm{Dom}(\phi) = \{\mu \in \mathcal{P}_2,\, \phi(\mu) < \infty\}$. The subdifferential of ϕ was defined in the Fréchet sense, see, e.g., Definition 10.1.1 of Ambrosio et al., 2005. We recall the definition of strong subdifferential as below.

Definition 2 (Strong subdifferential). *Given $\mu \in \mathcal{P}_2$, a vector field $\xi \in L^2(\mu)$ is a strong (Fréchet) subdifferential of ϕ at μ if for $v \in L^2(\mu)$,*

$$\phi((\mathrm{I_d} + v)_\# \mu) - \phi(\mu) \geqslant \langle \xi, v \rangle_\mu + o(\|v\|_\mu).$$

We denote by $\partial_{\mathcal{W}_2} \phi(\mu)$ the set of strong Fréchet subdifferentials of ϕ at μ (which may be empty).

There can be different ways to introduce convexity of functions on $\mathcal{P}_2$. The most common way is the convexity along geodesics, also known as "displacement convexity." In our analysis, we technically need the notation of convexity *along generalized geodesics* (a.g.g.), which is stronger than geodesic convexity. In short, displacement convexity is along the

geodesic from μ_1 to μ_2, which (in the simple case where there is a unique OT map T_1^2 from μ_1 to μ_2) is defined using interpolation $(1-t)\mathrm{I_d} + tT_1^2$ for $t \in [0,1]$. In contrast, convexity a.g.g. involves a third distribution ν and is defined using interpolation of the two OT maps from ν to μ_1 and μ_2 respectively.

Specifically, let $\nu \in \mathcal{P}_2^r$, $\mu_i \in \mathcal{P}_2$, $i = 1, 2$, and let T_ν^i be the OT map from ν to μ_i respectively. A *general geodesic* joining μ_1 to μ_2 (with base ν) is a curve of type

$$\mu_t^{1\to 2} := ((1-t)T_\nu^1 + tT_\nu^2)_\# \nu, \quad t \in [0,1]. \tag{4.11}$$

Definition 3 (Convexity a.g.g. (along generalized geodesics))**.** *For $\lambda \geqslant 0$, a functional ϕ on $\mathcal{P}_2$ is said to be λ-convex along generalized geodesics (a.g.g.) if for any $\nu \in \mathcal{P}_2^r$ and $\mu_1, \mu_2 \in \mathcal{P}_2$,*

$$\phi(\mu_t^{1\to 2}) \leqslant (1-t)\phi(\mu_1) + t\phi(\mu_2) - \frac{\lambda}{2}t(1-t)\mathcal{W}_\nu^2(\mu_1,\mu_2), \quad \forall t \in [0,1], \tag{4.12}$$

where $\mu_t^{1\to 2}$ is as in (4.11) *and*

$$\mathcal{W}_\nu^2(\mu_1,\mu_2) := \int_{\mathbb{R}^d} \|T_\nu^1(x) - T_\nu^2(x)\|^2 d\nu(x) \geqslant \mathcal{W}_2^2(\mu_1,\mu_2). \tag{4.13}$$

Note that the definition implies the following property which is useful in our analysis

$$\phi(\mu_t^{1\to 2}) \leqslant (1-t)\phi(\mu_1) + t\phi(\mu_2) - \frac{\lambda}{2}t(1-t)\mathcal{W}_2^2(\mu_1,\mu_2), \quad \forall t \in [0,1]. \tag{4.14}$$

The definition of convexity a.g.g. in Ambrosio et al., 2005, Section 9.2 is for the more general case when ν may not have density and the OT maps from ν to μ_i need to be replaced with optimal plans, and then the generalized geodesics may not be unique. In this paper, we only consider the case where ν has a density so we simplify the definition, see Ambrosio et al., 2005, Remark 9.2.3 (and make it slightly weaker but there is no harm for our purpose).

4.2.3 JKO scheme for Fokker-Planck equations

Consider the diffusion process (4.7) starting from $P \in \mathcal{P}_2$. It is known that under generic condition, as $t \to \infty$, ρ_t converges to the equilibrium distribution of (4.7) which has density

$$q \propto e^{-V}, \tag{4.15}$$

and the convergence is exponentially fast Bolley et al., 2012. The function V is called the *potential function* of q.

The evolution of ρ_t by FPE of the diffusion process can be interpreted as a continuous-time gradient flow under the $\mathcal{W}_2$-metric in the probability space $\mathcal{P}_2$. The JKO scheme (Jordan et al., 1998) computes a Wasserstein proximal GD which is a time discretization of the gradient flow. Specifically, define $G : \mathcal{P}_2^r \to \mathbb{R}$ as the KL-divergence w.r.t. q, i.e.,

$$\begin{aligned} G(\rho) &= \mathrm{KL}(\rho||q) = \mathcal{H}(\rho) + \mathcal{E}(\rho), \\ \mathcal{H}(\rho) &= \int \rho \log \rho, \quad \mathcal{E}(\rho) = c + \int V\rho, \end{aligned} \tag{4.16}$$

where c is a constant. More general G can be considered, see Section 4.4.1, and in this work we mainly focus on the case where G is the KL divergence as being considered in Jordan et al., 1998.

Under certain regularity condition of V, the JKO scheme computes a sequence of distributions ρ_n, $n = 0, 1, ...$, starting from $\rho_0 \in \mathcal{P}_2$. For a fixed step size $\gamma > 0$, and the scheme at the n-th step can be written as

$$\rho_{n+1} = \arg\min_{\rho \in \mathcal{P}_2} \Big(G(\rho) + \frac{1}{2\gamma} \mathcal{W}_2^2(\rho_n, \rho) \Big). \tag{4.17}$$

The scheme computes the $\mathcal{W}_2$-proximal Gradient Descent (GD) of G with step size γ, and can be equivalently written as

$$\rho_{n+1} = \mathrm{Prox}_{\gamma G}(\rho_n). \tag{4.18}$$

The original JKO paper Jordan et al., 1998 proved the convergence of the discrete-time solution $\{\rho_n\}$ (after interpolation over time) to the continuous-time solution ρ_t of the FPE (4.8) when step size $\gamma \to 0+$. In the context of flow-based generative models by neural networks, the discrete-time JKO scheme with finite γ was adopted and implemented as a flow network in Xu et al., 2023b. Our analysis in this work will prove the exponential convergence of ρ_n to q by the JKO scheme (including learning error), echoing the exponential convergence of the continuous-time dynamic (the FPE). This result leads to the guarantee

of generating data distributions up to (TV) error $O(\varepsilon)$ in $O(\log(1/\varepsilon))$ JKO steps. We will summarize the flow model and introduce needed theoretical assumptions in Section 4.3.

4.3 Setup of JKO flow model and assumptions

In this section, we summarize the mathematical setup for the JKO flow model and introduce the necessary theoretical assumptions for our analysis. The guarantee of generating the data distribution will be derived in Section 4.5 based on the exponential convergence of the $\mathcal{W}_2$-proximal GD (JKO scheme) in Section 4.4.

4.3.1 Forward and reverse processes of JKO flow model

The flow model implements an ODE model (transport equation), where both the forward process and the reverse process are computed by an invertible Residual Network Behrmann et al., 2019 or a neural-ODE network R. T. Chen et al., 2018; Grathwohl et al., 2018. The forward process consists of N steps, where each step is computed by a Residual Block - in the neural-ODE model, this is the neural ODE integration on a sub-time-interval $[t_n, t_{n+1}]$, and we also call it a Residual Block. The backward process consists of the N steps of the same flow network "backward in time," where each step computes the inverse map of the Residual Block, and in the neural-ODE model, this is via integrating the ODE in reverse time.

The forward and reverse process (without inversion error) are induced by a sequence of transport maps, T_n, $n = 1, \ldots, N$, which we will define more formally later. The two processes are summarized in (4.19),

$$\begin{aligned} &\text{(forward)} \quad p = p_0 \xrightarrow{T_1} p_1 \xrightarrow{T_2} \cdots \xrightarrow{T_N} p_N \approx q, \\ &\text{(reverse)} \quad p \approx q_0 \xleftarrow{T_1^{-1}} q_1 \xleftarrow{T_2^{-1}} \cdots \xleftarrow{T_N^{-1}} q_N = q. \end{aligned} \tag{4.19}$$

where p is the density of data distribution (when exists, otherwise a smoothified density by a short time diffusion), and q is the equilibrium density, typically chosen as Gaussian. Inversion error in the reverse process is considered in Section 4.3.3.

Forward process. In the forward process, the algorithm learns a sequence of T_n which transports from data distribution P to the equilibrium distribution Q which is typically the normal distribution. We denote the by q the density of Q, typically $\mathcal{N}(0, I)$, and p the density of the data distribution P when there is a one.

Following the neural-ODE framework used in Xu et al., 2023b, each step computes a transport map $T_{n+1} : \mathbb{R}^d \to \mathbb{R}^d$ which is the solution map of the ODE from time t_n to t_{n+1}, i.e.,

$$T_{n+1}(x_n) = x(t_{n+1}), \quad \text{where } x(t) \text{ solves } \dot{x}(t) = \hat{v}(x(t), t) \text{ on } [t_n, t_{n+1}],\ x(t_n) = x_n, \tag{4.20}$$

and $\hat{v}(x, t)$ is the velocity field on $\mathbb{R}^d$ parametrized by the n-th Residual Block. Equivalently, we have

$$T_{n+1}(x_n) = x_n + \int_{t_n}^{t_{n+1}} \hat{v}(x(t), t)dt, \quad x(t_n) = x_n. \tag{4.21}$$

In the implementation of the JKO scheme in a flow network, the learning of the N Residual Blocks is conducted progressively for $n = 1, \cdots, N$ by minimizing a training objective per step Xu et al., 2023b. We emphasize that, unlike other normalizing flow or diffusion models which are trained end-to-end, the training procedure here is done step-wise and progressively over the N Residual Blocks.

Once T_{n+1} is learned, it pushes from p_n to p_{n+1}, i.e.,

$$p_{n+1} = (T_{n+1})_{\#} p_n. \tag{4.22}$$

The sequence of transports starts from $p_0 = p$, which is the data density. Strictly speaking, we will introduce an initial short-time diffusion that smoothifies the data density p into ρ_δ which we set to be p_0 (see more in Section 4.5.1). The learning aims that after N steps the final p_N is close to the equilibrium density q.

Reverse process (without inversion error). The reverse process computes the inverse of the N-steps transport by inverting each T_n in the forward process. We first assume that T_n can be exactly inverted in computation which allows a simplified analysis. In practice, T_n^{-1}

can be implemented by fixed-point iteration Behrmann et al., 2019 or reverse-time ODE integration Grathwohl et al., 2018. The case when the inverse cannot be exactly computed is discussed in Section 4.3.3, where we need additional assumptions on the closeness of the computed inverse to the true inverse of T_n for our analysis.

The reverse process outputs generated samples, which are aimed to be close in distribution to the data samples, by drawing samples from q and pushing them through the reverse N steps. In terms of the sequence of probability densities generated by the process, the reverse process computes

$$q_n = (T_{n+1}^{-1})_{\#} q_{n+1}, \tag{4.23}$$

starting from $q_N = q$ and the output density is q_0. Theoretically, the data processing inequality for the KL-divergence applied to invertible transforms (Lemma 9) guarantees that if p_N is close to $q_N = q$, then q_0 is close to p_0, which is the data density (possibly after short-time smoothing). This allows us to prove the guarantee of $q_0 \approx p_0$ once we can prove that of $p_N \approx q$, the latter following the convergence of the $\mathcal{W}_2$-proximal GD up to a hopefully small learning error to be detailed below.

4.3.2 Learning assumptions of the forward process

We consider the sequence of densities p_n in the forward process in (4.19). Recall from Section 4.2.3 that for fixed step-size $\gamma > 0$, the n-th step classical JKO scheme finds p_{n+1} by minimizing

$$\min_{\rho \in \mathcal{P}_2} F_{n+1}(\rho) := G(\rho) + \frac{1}{2\gamma} \mathcal{W}_2^2(p_n, \rho), \tag{4.24}$$

where $G(\rho) = \mathrm{KL}(\rho || q)$. The learning in the n-th Residual Block in a JKO flow network computes the minimization via parametrizing the transport T_{n+1}. Here we briefly review the rational of solving (4.24) by solving for T_{n+1}, which leads to our assumption of the learned forward process.

JKO step by learning the transport. In the right hand side of (4.24), when both p_n and ρ are in $\mathcal{P}_2^r$, the Brenier Theorem (Theorem 1) implies the existence of a unique OT map T

from p_n to ρ. Consider the following minimization over the transport T,

$$\min_{T:\mathbb{R}^d\to\mathbb{R}^d} G(T_\# p_n) + \frac{1}{2\gamma}\mathbb{E}_{x\sim p_n}\|x - T(x)\|^2. \tag{4.25}$$

The following lemma, proved in section 4.6, shows that the minimizer T makes $T_\# p_n \in \mathcal{P}_2^r$:

Lemma 4. *Suppose $p_n \in \mathcal{P}_2^r$ makes $G(p_n) < \infty$, and T is a minimizer of* (4.25)*, then $T_\# p_n \in \mathcal{P}_2^r$.*

Thus, in (4.25) it is equivalent to minimize over T that renders $T_\# p_n \in \mathcal{P}_2^r$ and this means that the minimizer T is the OT map. One can also verify that (4.25) is equivalent to (4.24) in the sense that a minimizer T^* of (4.25) makes $(T^*)_\# p_n$ a minimizer of (4.24), and for a minimizer ρ^* of (4.24) the OT map from p_n to ρ^* is a minimizer of (4.25) Xu et al., 2023b, Lemma A.1.

Learning error in JKO flow network. In the n-step of the JKO flow network, the transport T in (4.25) is parameterized by a Residual block, and the learning cannot find the T that exactly minimizes (4.25) (and equivalently (4.24)) for three reasons:

(i) Approximation error: The minimization of (4.25) is over T constrained inside some neural network family $\mathcal{T}_\Theta$. When the function family $\mathcal{T}_\Theta$ is large enough to express the desired optimal transport from p_n to $\mathrm{Prox}_{\gamma G}(p_n)$, the solution can approximate the exact minimizer of (4.24), but this usually cannot be guaranteed.

(ii) Finite-sample effect: The training is computed on empirical data samples, while in this analysis, we focus on the minimization of population loss.

(iii) Imperfect optimization: The learning of neural networks is a non-convex optimization typically implemented by Stochastic Gradient Descent (SGD) over mini-batches and there is no guarantee of achieving a minimizer of the empirical loss.

As a result, the learned transport T_{n+1} finds a $p_{n+1} = (T_{n+1})_\# p_n$ that at most approximately minimizes (4.24). While the learned T_{n+1} is usually not the exact minimizer, we assume that it is regular enough such that p_{n+1} is still in $\mathcal{P}_2^r$, which would hold if, e.g., T_{n+1} has finite global Lipschitz constant by Lemma 5 (proved in section 4.6).

Lemma 5 (Global Lipschitz T transports from $\mathcal{P}_2^r$ to $\mathcal{P}_2^r$)**.** *Suppose $T : \mathbb{R}^d \to \mathbb{R}^d$ has a finite Lipschitz constant on $\mathbb{R}^d$, and $p \in \mathcal{P}_2^r$, then $T_\# p$ is also in $\mathcal{P}_2^r$.*

The finite Lipschitz constant of T_{n+1} (and its inverse) will be explained in Section 4.5.2. When p_n and p_{n+1} are both in $\mathcal{P}_2^r$, we have a unique invertible OT map from p_n to p_{n+1} by Brenier Theorem, and its (a.e.) inverse is the OT map from p_{n+1} to p_n. Specifically, we define

$$\begin{aligned} &T_n^{n+1} \text{ is the OT map from } p_n \text{ to } p_{n+1}, \quad p_n\text{-a.e.}, \\ &T_{n+1}^n \text{ is the OT map from } p_{n+1} \text{ to } p_n, \quad p_{n+1}\text{-a.e.}, \end{aligned} \tag{4.26}$$

and we have $T_{n+1}^n \circ T_n^{n+1} = \mathrm{I_d}$ p_n-a.e., $T_n^{n+1} \circ T_{n+1}^n = \mathrm{I_d}$ p_{n+1}-a.e.

Assumption on approximate first-order condition. For our analysis, we theoretically characterize the error in learning T_{n+1} by quantifying the error in the first-order condition. Specifically, the $\mathcal{W}_2$-gradient (strictly speaking, a sub-differential) of F_{n+1} at ρ can be identified as

$$\nabla_{\mathcal{W}_2} F_{n+1}(\rho) = \nabla_{\mathcal{W}_2} G(\rho) - \frac{T_\rho^n - \mathrm{I_d}}{\gamma}, \quad \rho\text{-a.e.} \tag{4.27}$$

where T_ρ^n is the OT map from ρ to p_n. Here we use $\nabla_{\mathcal{W}_2}\phi$ to denote the sub-differential $\partial_{\mathcal{W}_2}\phi$ assuming unique existence to simplify exhibition. The formal statement in terms of subdifferential is provided in Lemma 14 (which follows the argument of Ambrosio et al., 2005, Lemma 10.1.2) for more general G (which includes KL-divergence as a special case). For KL-divergence G, if V is differentiable, the sub-differential $\partial_{\mathcal{W}_2} G$ is reduced to the unique $\mathcal{W}_2$-gradient written as

$$\nabla_{\mathcal{W}_2} G(\rho) = \nabla V + \nabla \log \rho, \tag{4.28}$$

when $\rho \in \mathcal{P}_2^r$ has a well-defined score function.

If ρ is the exact minimizer of (4.24), we will have $\nabla_{\mathcal{W}_2} F_{n+1}(\rho) = 0$ (and for sub-differential the condition is $0 \in \partial_{\mathcal{W}_2} F_{n+1}(\rho)$). At p_{n+1} which is pushed-forward by the learned T_{n+1}, we denote the $\mathcal{W}_2$-gradient of F_{n+1} by the following (recall the definition of

T_{n+1}^n as in (4.26))

$$\xi_{n+1} := \nabla_{\mathcal{W}_2} F_{n+1}(p_{n+1}) = \nabla_{\mathcal{W}_2} G(p_{n+1}) - \frac{T_{n+1}^n - \mathrm{I_d}}{\gamma}, \quad p_{n+1}\text{-a.e.} \tag{4.29}$$

and it is interpreted as sub-differential when needed. While p_{n+1} differs from the exact minimizer, we assume it is close enough such that there is a subdifferential ξ_{n+1} that is small. In practice, the SGD algorithm to minimize the training objective (4.25) (assuming $\mathcal{T}_\Theta$ is expressive enough to approximate the exact minimizer T^*) would almost converge when the $\mathcal{W}_2$-gradient vector field ξ_{n+1} evaluated on data samples collectively give a small magnitude. We characterize this by a small $L^2(p_{n+1})$ norm of the (sub-)gradient ξ_{n+1} in our theoretical assumption. The assumptions on the learned transport T_{n+1} are summarized as follows:

Assumption 1 (Approximate n-th step solution). The learned transport T_{n+1} is invertible and both T_{n+1} and T_{n+1}^{-1} have finite Lipschitz constants on $\mathbb{R}^d$. In addition, for some $\varepsilon > 0$, $\exists \xi_{n+1} \in \partial_{\mathcal{W}_2} F_{n+1}(p_{n+1})$ s.t.

$$\|\xi_{n+1}\|_{p_{n+1}} \leqslant \varepsilon. \tag{4.30}$$

The error magnitude ε can be viewed as an algorithmic parameter that controls the accuracy of first-order methods, and similar assumptions have been made in the analysis of stochastic (noisy) gradient descent in vector space, see, e.g., Nemirovski et al., 2009; Nemirovsky and Yudin, 1983. We emphasize that theoretically ε does not need to be small but will enter the final error bound. Under Assumption 1, since T_{n+1} has a finite Lipschitz constant on $\mathbb{R}^d$, from $p_n \in \mathcal{P}_2^r$, $p_{n+1} = (T_{n+1})_\# p_n$ is also in $\mathcal{P}_2^r$ by Lemma 5. Then the subdifferential $\partial_{\mathcal{W}_2} F_{n+1}$ can be defined at p_{n+1} and characterized by Lemma 14.

Remark 4.3.1 (Global Lipschitzness). We require the global Lipschitzness of T_{n+1} and its inverse only to ensure that $p_{n+1} \in \mathcal{P}_2^r$ and data processing inequality in both directions (Lemma 9) can be applied. In particular, the Lipchitz constants do not enter the quantitative convergence bound in Section 4.5.1, when there is no inversion error. When there is an inversion error, the Lipschitz constant will theoretically enter the error bound, see more in

Section 4.5.2.

Remark 4.3.2 (T_{n+1} and T_n^{n+1}). Recall that T_{n+1} is the learned transport map and T_n^{n+1} is the OT map. In our setting (of imperfect minimization in the n-th step), both T_{n+1} and T_n^{n+1} push p_n to p_{n+1} but they are not necessarily the same. The notion of T_n^{n+1} is introduced only for theoretical purposes (the existence and invertibility are by Brenier Theorem), and there is no need for the learned map T_{n+1} to equal T_n^{n+1} for the theoretical result to hold.

4.3.3 Reverse process with inversion error

Considering potential inversion error in the reverse process, we denote the sequence of transports as S_n and the transported densities as $\tilde{q}_n$, that is,

$$\tilde{q}_n = (S_{n+1})_{\#}\tilde{q}_{n+1}, \tag{4.31}$$

from $\tilde{q}_N = q_N = q$. The reverse process with and without inversion error is summarized as

$$\begin{aligned} &\text{(exact reverse)} \qquad q_0 \xleftarrow{T_1^{-1}} q_1 \xleftarrow{T_2^{-1}} \cdots \xleftarrow{T_N^{-1}} q_N = q, \\ &\text{(computed reverse)} \qquad \tilde{q}_0 \xleftarrow{S_1} \tilde{q}_1 \xleftarrow{S_2} \cdots \xleftarrow{S_N} \tilde{q}_N = q. \end{aligned} \tag{4.32}$$

The computed transport S_n is not the same as T_n^{-1} but the algorithm aims to make the inversion error small. For our theoretical analysis, we make the following assumption on the error.

Assumption 2 (Inversion error). For $n = N, \cdots, 1$, the computed reverse transport S_n satisfies that

$$\|T_n \circ S_n - \mathrm{I_d}\|_{\tilde{q}_n} \leqslant \varepsilon_{\mathrm{inv}}. \tag{4.33}$$

The quantity $\|T_n \circ S_n - \mathrm{I_d}\|_{\tilde{q}_n}^2$ can be empirically estimated by sample average, namely the mean-squared error

$$\mathrm{MSE}_{\mathrm{inv}} = \frac{1}{n_{\mathrm{inv}}} \sum_i \|T_n \circ S_n(x_i) - x_i\|^2, \quad x_i = S_{n+1} \circ \cdots \circ S_{N-1}(z_i), \quad z_i \sim q,$$

computed from n_{inv} test samples. With sufficiently large n_{inv}, one can use $\mathrm{MSE}_{\mathrm{inv}}$ to monitor the inversion error of the reverse process and enhance numerical accuracy when needed.

It was empirically shown in Xu et al., 2023b that the inversion error computed on testing samples (though all the N blocks) can be made small towards the floating-point precision in the neural-ODE model, and keeping the inversion error small is crucial for the success of the JKO flow network, which can also be seen from our theoretical results.

4.4 Convergence of forward process

The current paper mainly concerns the application to flow generative networks where G is the KL divergence. In this section, we prove the exponentially fast convergence of the forward process which applies to potentially a more general class of G as long as the subdifferential calculus in $\mathcal{P}_2$ can be conducted (see Section 10.1 of Ambrosio et al., 2005) and G is strongly convex along generalized geodesics (a.g.g., see Definition 3), which may be of independent interest.

We will revisit the KL-divergence G as a special case and prove the generation guarantee of the reverse process in Section 4.5. All proofs and technical lemmas are provided in Appendix 4.7.

4.4.1 Conditions on G and V

We introduce the more general condition of G needed by the forward process convergence.

Assumption 3 (General condition of G). $G : \mathcal{P}_2 \to (-\infty, +\infty]$ is lower semi-continuous, $\mathrm{Dom}(G) \subset \mathcal{P}_2^r$; G is λ-convex a.g.g. in $\mathcal{P}_2$.

The first part of Assumption 3 ensures that the strong subdifferential $\partial_{\mathcal{W}_2} G(\rho)$ can be defined, see Definition 2. The strong convexity of G is used to prove the exponential convergence of the (approximated) $\mathcal{W}_2$-proximimal Gradient Descent in the forward process.

Next, we show that the KL divergence G satisfies the general condition under certain general conditions of the potential function V plus its strong convexity. We also introduce an upper bound of λ due to a rescaling argument.

KL divergence G**.** Recall that the KL-divergence $G(\rho) = \mathcal{H}(\rho) + \mathcal{E}(\rho)$ as defined in (4.16), where $\mathcal{E}(\rho) = \int V\rho$ involves the potential function V of the equilibrium density. We introduce the following assumption on V:

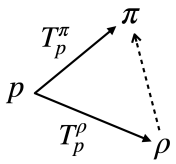
Assumption 4 (Condition of V the potential of q). The potential function $V : \mathbb{R}^d \to (-\infty, \infty]$ is proper, lower semi-continuous, and $V^- := \max\{-V, 0\}$ is bounded; $V(x)$ is λ-strongly convex on $\mathbb{R}^d$, and $q \propto e^{-V}$ is in $\mathcal{P}_2^r$.

The first part of Assumption 4 is for the sub-differential calculus of $\mathcal{E}(\rho) = \int V\rho$ in $\mathcal{P}_2$. The λ-strong convexity is used to make $\mathcal{E}(\rho)$ (and subsequently $G(\rho)$) λ-convex a.g.g. Under such condition of V, the KL divergence $G(\rho)$ satisfies Assumption 3, which is verified in Lemma 17. Thus our result applies to the KL divergence as being used in the JKO flow model. For the important case when q is standard normal, $V(x) = \|x\|^2/2$, and $\lambda = 1$.

Positive and bounded λ**.** Note that for strongly convex V we can use a scaling argument to make λ bounded to be $O(1)$ without loss of generality. Specifically, for V that is λ-convex on $\mathbb{R}^d$, the function $x \mapsto V(ax)$ for $a > 0$ is $(a^2\lambda)$-convex. This means that for q that has a strongly convex V as the potential function, one can rescale samples from q to make V strongly convex with $\lambda \leqslant 1$. In the case where G is KL divergence, the λ-convexity of G has the same λ as that of V. Thus, for the general G we assume its λ is also bounded by 1.

Assumption 5 (λ bounded). In Assumptions 3 and 4, $0 < \lambda \leqslant 1$.

Our technique can potentially extend to analyze the $\lambda = 0$ case, where an algebraic $O(1/n)$ convergence rate is expected instead of the exponential rate proved in Theorem 8. For the application to flow-based generative model, one would need the equilibrium density $q \propto e^{-V}$ convenient to sample from, and thus the normal density (corresponding to $V(x) = \|x\|^2/2$) is the most common choice and the other choices usually render $\lambda > 0$ (to enable fast sampling of the starting distribution). We thus leave the $\lambda = 0$ case to future work.



$$p, \pi, \rho \in \mathcal{P}_2^r, G \text{ is } \lambda\text{-convex a.g.g.,}$$
$$G(\pi) - G(\rho) \geqslant \left\langle \nabla_{\mathcal{W}_2} G(\rho) \circ T_p^\rho, T_p^\pi - T_p^\rho \right\rangle_p + \tfrac{\lambda}{2}\mathcal{W}_2^2(\pi, \rho)$$

$$\pi, \rho \in \mathbb{R}^d, g \text{ is } \lambda\text{-convex,}$$
$$g(\pi) - g(\rho) \geqslant \langle \nabla g(\rho), \pi - \rho \rangle + \tfrac{\lambda}{2}\|\pi - \rho\|^2$$

FIGURE 4.2: The monotonicity of a.g.g.-convex G in $\mathcal{P}_2$ proved in Lemma 6. The dotted line indicates the general geodesic between ρ and π.

4.4.2 EVI and convergence of the forward process

The a.g.g. λ-convexity of G leads to the following lemma, which is important for our analysis.

Lemma 6 (Monotonicity of G)**.** *Let $p, \rho \in \mathcal{P}_2^r$, $\pi \in \mathcal{P}_2$, and denote by T_p^ρ and T_p^π the OT maps from p to ρ and to π respectively. Suppose G satisfies Assumption 3, then for any $\eta \in \partial_{\mathcal{W}_2} G(\rho)$,*

$$G(\pi) - G(\rho) \geqslant \left\langle \eta \circ T_p^\rho, T_p^\pi - T_p^\rho \right\rangle_p + \frac{\lambda}{2}\mathcal{W}_2^2(\pi, \rho).$$

The relationship among p, ρ, π is illustrated in Figure 4.2, which also includes an analog to the strong-convex function in Euclidean space. This lemma extends Lemma 4 in Salim et al., 2020 and originally the argument in Section 10.1.1.B of Ambrosio et al., 2005. We include a proof in section 4.7 for completeness.

Based on the monotonicity lemma and the condition of small strong subdifferential ξ_{n+1} in Assumption 1, we are ready to drive the discrete-time *Evolution Variational Inequality* (EVI) Ambrosio et al., 2005, Chapter 4 for the (approximate) JKO scheme.

Lemma 7 (EVI for approximate JKO step)**.** *Given $\pi \in \mathcal{P}_2$, suppose G satisfies Assumption 3 with $\lambda \in (0, 1]$, and $0 < \gamma < 2$. If $p_0 \in \mathcal{P}_2^r$, and Assumption 1 holds for $n = 0, 1, \cdots$, then for all n,*

$$\left(1 + \frac{\gamma\lambda}{2}\right)\mathcal{W}_2^2(p_{n+1}, \pi) + 2\gamma\left(G(p_{n+1}) - G(\pi)\right) \leqslant \mathcal{W}_2^2(p_n, \pi) + \frac{2\gamma}{\lambda}\varepsilon^2. \tag{4.34}$$

Technically, we require that the step size $\gamma < 2$ (and the proof shows that 2 can be changed to another $\gamma_{\max} > 1$ which affects the constant in the final bound and does not affect the order). The rationale is as follows: First, it has been empirically observed that

successful computation of the JKO flow model in practice needs the step size not to exceed a certain maximum value, which is an algorithmic parameter Xu et al., 2023b. Setting the step size too large may lead to difficulty in training the Residual blocks as well as in maintaining the inversion error small. Meanwhile, from the formulation of the JKO scheme, it can be seen that for large γ, the proximal GD in (4.17) approaches the global minimization of $G(\rho)$, which asks for the flow to transport from the current density to the target density q in one step. Though the proximal GD (as a backward Euler scheme) does not impose a step-size constraint, the optimization problem (4.17) is in principle easier with a small (but no need to converge to zero) step size. We thus adopt the $\gamma < 2$ condition here, which is motivated by practice and for exhibition simplicity.

The EVI directly leads to the N-step convergence of the forward process, which achieves $O(\varepsilon)$ $\mathcal{W}_2$-error and $O(\varepsilon^2)$ gap from the optimal objective value in $N \lesssim \log(1/\varepsilon)$ JKO steps.

Theorem 8 (Convergence of forward process)**.** *Suppose $q \in \mathcal{P}_2$ is the global minimum of G, and the other assumptions are the same as Lemma 7, then*

$$\mathcal{W}_2^2(p_n, q) \leqslant \left(1 + \frac{\gamma\lambda}{2}\right)^{-n} \mathcal{W}_2^2(p_0, q) + \frac{4\varepsilon^2}{\lambda^2}, \quad n = 1, 2, \cdots \tag{4.35}$$

In particular, if

$$n \geqslant \frac{8}{\gamma\lambda} \left(\log \mathcal{W}_2(p_0, q) + \log(\lambda/\varepsilon)\right), \tag{4.36}$$

then

$$\mathcal{W}_2(p_n, q) \leqslant \sqrt{5}\frac{\varepsilon}{\lambda} \quad \textit{and} \quad G(p_{n+1}) - G(q) \leqslant \frac{9}{2\gamma}\left(\frac{\varepsilon}{\lambda}\right)^2. \tag{4.37}$$

Remark 4.4.1 (Comparison to Salim et al., 2020). The convergence rates of Wasserstein proximal GD were previously studied in Salim et al., 2020, and our proof techniques, namely the monotonicity of G plus discrete-time EVI, are similar to the analysis therein. However, the setups differ in several aspects: first, we consider the "fully-backward" proximal GD, i.e., the JKO scheme, while Salim et al., 2020 focuses on the forward-backward scheme to minimize G having a decomposed form $\mathcal{E} + \mathcal{H}$ (which does cover the KL convergence as

a special case). Second, Salim et al., 2020 assumed the exact solution of the proximal step while our analysis takes into account the error ε in the first-order condition Assumption 1, which is more realistic for neural network-based learning. At last, Salim et al., 2020 assumed L-smoothness of V, namely ∇V is L-Lipschitz, and step size $\gamma < 1/L$, as a result of the forward step in the splitting scheme, which is not needed in our fully backward scheme. The motivation for Salim et al., 2020 is to understand the discretized Wasserstein gradient flow and to recover the same convergence rates as the (forward-backward) proximal GD in the vector space. Our analysis is motivated by the JKO flow network, and the forward process convergence is an intermediate result to prove the generation guarantee of the reverse process.

4.5 Generation guarantee of reverse process

In this section, we first consider the reverse process as in (4.19), called the *exact* reserves process, when there is no inversion error. In Section 4.5.1, we prove a KL (and TV) guarantee of generating by q_0 any data distribution P in $\mathcal{P}_2^r$, and extend to P with no density by introducing a short-time initial diffusion.

Taking into account the inversion error, we consider the sequence $\tilde{q}_n$ induced by S_n in (4.32), called the *computed* reverse process, where the inversion error satisfies Assumption 2. In Section 4.5.2, we prove a closeness bound of $\tilde{q}_0$ to q_0 in $\mathcal{W}_2$, which leads to a $\mathcal{W}_2$-KL mixed generation guarantee of $\tilde{q}_0$ based on the proved guarantee of q_0.

4.5.1 Convergence guarantee without inversion error

We start by presenting convergence analysis assuming there are no errors in the reverse process; then we extend to the more practical situation considering inversion errors.

KL (TV) guarantee of generating P with density

Consider the transport map over the N steps in (4.19) denoted as

$$T_1^N := T_N \circ \cdots \circ T_1. \tag{4.38}$$

Under Assumption 1, each T_n (and its inverse) is invertible and global Lipschitz on $\mathbb{R}^d$, and then so is T_1^N (and its inverse). We have

$$p_N = (T_1^N)_{\#} p_0, \quad q_N = (T_1^N)_{\#} q_0.$$

The following lemma follows from the data processing inequality of KL, which allows us to obtain KL bound of $p_0 = \rho_\delta$ and q_0 from that of p_N and $q_N = q$.

Lemma 9 (Bi-direction data processing inequality)**.** *If $T : \mathbb{R}^d \to \mathbb{R}^d$ is invertible and for two densities p and q on $\mathbb{R}^d$, $T_{\#}p$ and $T_{\#}q$ also have densities, then*

$$\mathrm{KL}(p||q) = \mathrm{KL}(T_{\#}p||T_{\#}q).$$

The following corollary establishes an $O(\varepsilon^2)$ KL bound in $N \lesssim \log(1/\varepsilon)$ JKO steps, which implies an $O(\varepsilon)$ TV bound by Pinsker's inequality.

Corollary 10 (KL guarantee for $P \in \mathcal{P}_2^r$)**.** *Suppose $G(\rho) = \mathrm{KL}(\rho||q)$, the potential function V satisfies Assumption 4 with $\lambda \in (0, 1]$, and $0 < \gamma < 2$. Suppose $P \in \mathcal{P}_2^r$ with density p, let $p_0 = p$, and Assumption 1 holds for some ε for all n. Then, let*

$$N = \left\lceil \frac{8}{\gamma\lambda} \left(\log \mathcal{W}_2(p_0, q) + \log(\lambda/\varepsilon) \right) \right\rceil, \tag{4.39}$$

the generated density q_0 of the reverse process satisfies that

$$\mathrm{KL}(p||q_0) \leqslant \frac{9}{2\gamma} \left(\frac{\varepsilon}{\lambda} \right)^2, \quad \mathrm{TV}(p, q_0) \leqslant \frac{3}{2\sqrt{\gamma}} \frac{\varepsilon}{\lambda}. \tag{4.40}$$

Guarantee of generating $P \in \mathcal{P}_2$ up to initial short diffusion

For $P \in \mathcal{P}_2$ that may not have a density, we first obtain $\rho_\delta \in \mathcal{P}_2^r$ that is close to P in $\mathcal{W}_2$ by a short-time initial diffusion (specifically, the OU process as introduced in Section 4.2.3) up to time $\delta > 0$, as shown in Lemma 18. The short-time initial diffusion was used in Lee et al., 2023 and called "early stopping" in H. Chen et al., 2023. It is also used in practice by flow model Xu et al., 2023b as well as score-based diffusion models to bypass the irregularity of data distribution Song, Sohl-Dickstein, et al., 2021. In principle, one can also use the

Brownian motion only (corresponding to convolving P with Gaussian kernel) to obtain ρ_δ. Here we use the OU process to stay in line with the literature.

The introduction of ρ_δ allows us to prove a guarantee of $\mathrm{KL}(\rho_\delta||q_0)$ in the following corollary, which is the same type of result as H. Chen et al., 2023, Theorem 2.

Corollary 11 (KL guarantee for $P \in \mathcal{P}_2$ from ρ_δ)**.** *Suppose $P \in \mathcal{P}_2$, and the conditions on G, V, λ and γ are the same as in Corollary 10. Then $\forall \varepsilon' > 0$,there exists $\delta > 0$ s.t. $\mathcal{W}_2(P, \rho_\delta) < \varepsilon'$ and, with $p_0 = \rho_\delta$ and Assumption 1 holds for some ε for all n, let N as in* (4.39)*, the generated density q_0 of the reverse process makes $\mathrm{KL}(\rho_\delta||q_0)$ and $TV(\rho_\delta, q_0)$ satisfy the same bounds as in* (4.40)*.*

The corollary shows that there can be a density $\rho_\delta \in \mathcal{P}_2^r$ that is arbitrarily close to P in $\mathcal{W}_2$, such that the output density q_0 of the reverse process can approximate ρ_δ up to the same error as in Corollary 10. Note that the corollary holds when the potential function V of q satisfies the general condition Assumption 4, and the OU process (corresponding to Gaussian q and quadratic V) is only used in constructing ρ_δ.

4.5.2 Convergence guarantee with inversion error

To prove the $\mathcal{W}_2$ control between $\tilde{q}_0$ and q_0, we first introduce a Lipschitz condition on the learned transport map T_n and the computed reverse transport map S_n.

Lipschitz constant of computed transport maps

Previously in Assumption 1, we required the learned transport map T_{n+1} to be invertible and globally Lipschitz on $\mathbb{R}^d$. Here, the additional assumption is summarized as follows:

Assumption 6 (Lipschitz condition on T_n^{-1} and S_n). There is $K > 0$ s.t. T_n^{-1} is Lipschitz on $\mathbb{R}^d$ with Lipschitz constant $e^{\gamma K}$ for all $n = N, \ldots, 1$; S_n has finite Lipschitz constant on $\mathbb{R}^d$ for all n.

The assumed Lipschitz constants are theoretical and motivated by neural ODE models, to be detailed below. Our analysis of $\mathcal{W}_2(\tilde{q}_0, q_0)$ applies to any type of flow network (like invertible ResNet) as long as Assumption 6 on T_n and S_n holds.

We justify the Lipschitz conditions in Assumptions 1 and 6 under the framework of

neural ODE flow, namely (4.20)(4.21), including the Lipschitz constant $e^{\gamma K}$ of T_n and inverse. Specifically, by the elementary Lemma 19 proved in Appendix 4.8.3, we know that if T_{n+1} can be numerically exactly computed as (4.21) and $\hat{v}(x,t)$ on $\mathbb{R}^d \times [t_n, t_{n+1}]$ satisfies a uniform x-Lipschitz condition with Lipschitz constant K, then both T_{n+1} and its inverse are Lipschitz on $\mathbb{R}^d$ with Lipschitz constant $e^{\gamma K}$. We will assume the same K throughout time for simplicity. In practice, Lipschitz regularization techniques can be applied to the neural network parametrized $\hat{v}(x,t)$, and the global Lipchitz bound of $\hat{v}$ on $\mathbb{R}^d$ can be achieved by "clipping" $\hat{v}$ to vanish outside some bounded domain of x. In addition, in a neural-ODE-based flow model, the reverse process is by integrating the neural ODE in reverse time, and thus we can also expect the same Lipschitz property of S_{n+1} (because $-\hat{v}(x,t)$ satisfies the same x-Lipschitz bound with $\hat{v}(x,t)$).

While the computed transport T_n and S_n often differ from the exact numerical integration of the ODE, we still expect the Lipschitz property to retain. For general flow models which may not be neural ODE, we impose the same theoretical assumption. At last, the global Lipschitz condition of the T_n and S_n may be theoretically relaxed by combining with truncation arguments of the probability distributions, which is postponed here.

$\mathcal{W}_2$-control of the computed reverse process from the exact one

Proposition 12. *Suppose $G(\rho) = \mathrm{KL}(\rho||q)$, the potential function V satisfies Assumption 4, the computed transports maps T_n and S_n satisfy Assumption 2 and Assumption 6. Then*

$$\mathcal{W}_2(\tilde{q}_0, q_0) \leqslant \frac{\varepsilon_{\mathrm{inv}}}{\gamma K} e^{\gamma K(N+1)}. \tag{4.41}$$

A continuous-time counterpart of Proposition 12 was derived in Albergo and Vanden-Eijnden, 2023, Proposition 3. We include a proof in Appendix 4.8.3 for completeness. The proof uses a coupling argument of the (discrete-time) ODE flow, which as has been pointed out in S. Chen, Chewi, et al., 2023, obtains a growing factor $e^{\gamma KN}$ in the $\mathcal{W}_2$-bound as shown in (4.41). To overcome this exponential factor, S. Chen, Chewi, et al., 2023 adopted an SDE corrector step. Here, without involving any corrector step, we show that the factor

$e^{\gamma KN}$ can be controlled at the order of some negative power of ε thanks to the exponential convergence in the forward process. This is because N can be chosen to be at the order of $\log(1/\varepsilon)$ as in (4.39), then $e^{\gamma KN}$ can be made $O(\varepsilon^{-\alpha})$ for some $\alpha > 0$. As a result, the $\mathcal{W}_2$-error (4.41) can be suppressed if ε_{inv} can be made smaller than a higher power of ε.

More specifically, combined with the analysis of the forward process, we arrive at the following.

Corollary 13 (Mixed bound with inversion error)**.** *Suppose $G(\rho) = \text{KL}(\rho||q)$, the potential function V satisfies Assumption 4 with $\lambda \in (0, 1]$, and $0 < \gamma < 2$. Suppose the computed transports maps T_n and S_n satisfy the Assumptions 1, 2, 6 for some ε and ε_{inv} for all n. Suppose $P \in \mathcal{P}_2^r$ with density p, let $p_0 = p$ and N as in* (4.39)*, then the generated density $\tilde{q}_0$ of the computed reverse process satisfies that*

$$\mathcal{W}_2(\tilde{q}_0, q_0) \leqslant \frac{e^{2\gamma K}}{\gamma K}(\mathcal{W}_2(p_0, q)\lambda)^{8K/\lambda}\frac{\varepsilon_{\text{inv}}}{\varepsilon^{8K/\lambda}}, \tag{4.42}$$

and q_0 satisfies the KL and TV bounds to p as in (4.40)*.*

The corollary implies that if ε_{inv} can be made small, then the $\mathcal{W}_2$ bound can be made equal to or smaller than ε in order. For example, if $\varepsilon_{\text{inv}} = O(\varepsilon^{8K/\lambda+1})$, then we have $\mathcal{W}_2(\tilde{q}_0, q_0) = O(\varepsilon)$. This suggests that if one focuses on getting $\text{KL}(p_N||q)$ small in the forward process, then maintaining an inversion error small is crucial for the generation quality of the flow model in the reverse process.

At last, when P is merely in $\mathcal{P}_2$ and does not have density, then one can start the forward process from $p_0 = \rho_\delta$ same as in Section 4.5.1. Then we have the same $\mathcal{W}_2$-bound between $\tilde{q}_0$ and q_0 as in (4.42), and q_0 is close to ρ_δ in the sense of Corollary 11.

4.6 Proofs and lemmas in Section 4.3

4.6.1 Lemma on the $\mathcal{W}_2$-(sub)gradient

Lemma 14. *Suppose $G : \mathcal{P}_2(X) \to (-\infty, +\infty]$ is lower semi-continuous and $\text{Dom}(G) \subset \mathcal{P}_2^r$. Let $\gamma > 0$, $p \in \mathcal{P}_2^r$, and*

$$F(\rho) = G(\rho) + \frac{1}{2\gamma}\mathcal{W}_2^2(p, \rho). \tag{4.43}$$

If (at $\rho \in \mathcal{P}_2^r$, $\partial_{\mathcal{W}_2} F(\rho)$ is non empty and) $\xi \in \partial_{\mathcal{W}_2} F(\rho)$, then

$$\xi + \frac{T_\rho^p - \mathrm{I_d}}{\gamma} \in \partial_{\mathcal{W}_2} G(\rho).$$

The argument follows that in Lemma 10.1.2 in Ambrosio et al., 2005, and we include a proof for completeness.

Proof of Lemma 14. We are to verify that

$$\eta := \xi + \frac{T_\rho^p - \mathrm{I_d}}{\gamma}$$

is a strong subdifferential of G at ρ. By Definition 2, it suffices to show that for any $v \in L^2(\rho)$ and $\delta \to 0$,

$$G((\mathrm{I_d} + \delta v)_{\#}\rho) - G(\rho) \geqslant \delta \langle \eta, v \rangle_\rho + o(\delta). \tag{4.44}$$

By construction,

$$\langle \eta, v \rangle_\rho = \langle \xi, v \rangle_\rho + \frac{1}{\gamma} \langle T_\rho^p - \mathrm{I_d}, v \rangle_\rho,$$

and since ξ is a strong subdifferential of F at ρ,

$$F((\mathrm{I_d} + \delta v)_{\#}\rho) - F(\rho) \geqslant \delta \langle \xi, v \rangle_\rho + o(\delta).$$

Combining the two and by the definition of F, we can deduce (4.44) as long as we can show that

$$\frac{1}{2}\mathcal{W}_2(p, (\mathrm{I_d} + \delta v)_{\#}\rho)^2 + o(\delta) \leqslant \frac{1}{2}\mathcal{W}_2(p, \rho)^2 - \langle T_\rho^p - \mathrm{I_d}, \delta v \rangle_\rho. \tag{4.45}$$

To show (4.45), note that by Brenier Theorem (ii),

$$\mathcal{W}_2(p, \rho)^2 = \int \|x - T_\rho^p(x)\|^2 \rho(x) dx = \|\mathrm{I_d} - T_\rho^p\|_\rho^2.$$

Thus,

$$\begin{aligned} \frac{1}{2}\mathcal{W}_2(p, \rho)^2 - \langle T_\rho^p - \mathrm{I_d}, \delta v \rangle_\rho &= \frac{1}{2}\|\mathrm{I_d} - T_\rho^p\|_\rho^2 + \langle \mathrm{I_d} - T_\rho^p, \delta v \rangle_\rho \\ &= \frac{1}{2}\|(\mathrm{I_d} + \delta v) - T_\rho^p\|_\rho^2 - \frac{1}{2}\|\delta v\|_\rho^2. \end{aligned} \tag{4.46}$$

Note that, because $v \in L^2(\rho)$,

$$\|\delta v\|_\rho^2 = O(\delta^2)$$

and

$$\|(\mathrm{I_d} + \delta v) - T_\rho^p\|_\rho^2 = \int_{\mathbb{R}^d} \|(\mathrm{I_d} + \delta v)(x) - T_\rho^p(x)\|^2 \rho(x) dx \geqslant \mathcal{W}_2((\mathrm{I_d} + \delta v)_{\#}\rho, p)^2.$$

Putting together, this gives that

$$(4.46) \geqslant \frac{1}{2}\mathcal{W}_2((\mathrm{I_d} + \delta v)_{\#}\rho, p)^2 + O(\delta^2)$$

which implies (4.45). □

4.6.2 Proofs of Lemma 4 and Lemma 5

Proof of Lemma 4. First, the minimizer makes the r.h.s. finite because $T = \mathrm{I_d}$ makes it finite: When T is identity, the r.h.s. equals $G(p_n) < \infty$. As a result, $\tilde{p} := T_{\#}p_n$ needs to have density because otherwise the KL divergence $G(\tilde{p}) = +\infty$.

It remains to show that $M_2(\tilde{p}) < \infty$. By definition,

$$\begin{aligned} M_2(\tilde{p}) &= \int_{\mathbb{R}^d} \|x\|^2 \tilde{p}(x) dx \\ &= \mathbb{E}_{x\sim p_n} \|T(x)\|^2 \\ &\leqslant 2(\mathbb{E}_{x\sim p_n}\|x\|^2 + \mathbb{E}_{x\sim p_n}\|x - T(x)\|^2), \end{aligned}$$

where $\mathbb{E}_{x\sim p_n}\|x\|^2 = M_2(p_n) < \infty$, and, at the minimizer T, $\mathbb{E}_{x\sim p_n}\|x - T(x)\|^2$ also needs to be finite due to that it is in the 2nd term of (4.25). □

Proof of Lemma 5. First, $T_{\#}p$ has a density (so the notation is well-defined) due to that T has a finite Lipschitz constant on $\mathbb{R}^d$. Suppose the Lipschitz constant is L. It remains to show that $M_2(T_{\#}p) < \infty$. By definition,

$$M_2(T_{\#}p) = \mathbb{E}_{x\sim p}\|T(x)\|^2,$$

and we have that $\forall x \in \mathbb{R}^d$,

$$\|T(x)\| \leqslant \|T(0)\| + \|T(x) - T(0)\| \leqslant \|T(0)\| + L\|x\|.$$

Thus,

$$\mathbb{E}_{x\sim p}\|T(x)\|^2 \leqslant 2(\|T(0)\|^2 + L^2\mathbb{E}_{x\sim p}\|x\|^2) = 2(\|T(0)\|^2 + L^2 M_2(p)) < \infty,$$

which shows that $M_2(T_\# p) < \infty$. □

4.7 Proofs and lemmas in Section 4.4

4.7.1 Technical lemmas in Section 4.4.1

Lemma 15. *$\mathcal{H}(\rho)$ is convex a.g.g. in $\mathcal{P}_2$.*

Proof. The a.g.g.-convexity of functional in the form of $\mathcal{F}(\rho) = \int F(\rho(x))dx$ in $\mathcal{P}_2$ is established in Proposition 9.3.9 of Ambrosio et al., 2005 when $F : [0, +\infty) \to (-\infty, \infty]$ is a proper, lower semi-continuous convex function satisfying that $s \mapsto s^d F(s^{-d})$ is convex and non-increasing on $(0, +\infty)$. The entropy $\mathcal{H}(\rho) = \mathcal{F}(\rho)$ with $F(s) = s\log s$, and this F satisfies the above conditions. □

Lemma 16. *Under Assumption 4, $\mathcal{E}(\rho)$ is λ-convex a.g.g. in $\mathcal{P}_2$.*

Proof. This is a direct result of Proposition 9.3.2(i) of Ambrosio et al., 2005, noting that assuming the boundedness of V^- implies the growth condition needed in Section 9.3 therein. The proof of Proposition 9.3.2(i) shows that $\mathcal{E}(\rho)$ is λ-convex along any interpolation curve which implies λ-convexity a.g.g. □

Lemma 17. *Under Assumptions 4, the KL divergence $G(\rho)$ defined in* (4.16) *satisfies Assumption 3.*

Proof. The lower semi-continuity follows from that of $\mathcal{H}(\rho)$ and the condition on V in Assumption 4. The domain of G is restricted to ρ with density because $\mathcal{H}(\rho)$ diverges otherwise. The a.g.g. λ-convexity of G directly follows from Lemma 15 and Lemma 16. □

4.7.2 Proofs in Section 4.4.2

Proof of Lemma 6. The unique existences of T_p^ρ and T_p^π are by Brenier Theorem. Since $\rho \in \mathcal{P}_2^r$, the map T_p^ρ has an inverse denoted by T_ρ^p which is defined ρ-a.e. Under Assumption 3 first part, the strong subdifferential of $\partial_{\mathcal{W}_2} G(\rho)$ is well-defined and we assume η is one of them.

Let $v := T_p^\pi \circ T_\rho^p - \mathrm{I_d}$. One can verify that $v \in L^2(\rho)$, since $\|T_p^\pi \circ T_\rho^p\|_\rho^2 = M_2(\pi)$, $\|\mathrm{I_d}\|_\rho^2 = M_2(\rho)$, and both are finite. By definition, for $\delta \in [0,1]$,

$$(\mathrm{I_d} + \delta v)_\# \rho = (1-\delta)\rho + \delta\pi = (T_p^\rho + \delta(T_p^\pi - T_p^\rho))_\# p. \tag{4.47}$$

We also have

$$\langle \eta, v \rangle_\rho = \left\langle \eta \circ T_p^\rho, T_p^\pi - T_p^\rho \right\rangle_p. \tag{4.48}$$

Since $v \in L^2(\rho)$, by that $\eta \in \partial_{\mathcal{W}_2} G(\rho)$ and the definition of strong subdifferential (Definition 2), with $\delta \to 0+$ we have

$$G\left((\mathrm{I_d} + \delta v)_\# \rho\right) \geqslant G(\rho) + \delta \langle \eta, v \rangle_\rho + o(\delta). \tag{4.49}$$

Combined with (4.47)(4.48), this gives

$$G\left((T_p^\rho + \delta(T_p^\pi - T_p^\rho))_\# p\right) - G(\rho) \geqslant \delta \left\langle \eta \circ T_p^\rho, T_p^\pi - T_p^\rho \right\rangle_p + o(\delta). \tag{4.50}$$

Meanwhile, by the λ-convexity of G a.g.g. (Definition 3), and specifically (4.14), we have

$$G\left((T_p^\rho + \delta(T_p^\pi - T_p^\rho))_\# p\right) \leqslant (1-\delta)G(\rho) + \delta G(\pi) - \frac{\lambda}{2}\delta(1-\delta)\mathcal{W}_2(\rho,\pi)^2. \tag{4.51}$$

Comparing (4.50) and (4.51), we have

$$G(\pi) - G(\rho) \geqslant \left\langle \eta \circ T_p^\rho, T_p^\pi - T_p^\rho \right\rangle_p + \frac{\lambda}{2}(1-\delta)\mathcal{W}_2(\rho,\pi)^2 + o(1).$$

We get the conclusion by letting $\delta \to 0+$. □

Proof of Lemma 7. By Assumption 1, in the n-th step, the learned forward process transport T_{n+1} pushes from p_n to p_{n+1}, and both are in $\mathcal{P}_2^r$. This holds for $n = 0, \cdots, N-1$, where $p_0 \in \mathcal{P}_2^r$ is by the lemma assumption. By the Brenier Theorem, the OT map from p_n to p_{n+1} is denoted as T_n^{n+1}, which is uniquely defined p_n-a.e. Let T_{n+1}^n be the OT map from p_{n+1} to p_n, and it is also the p_{n+1}-a.e. inverse of T_n^{n+1}. We use the short-hand notation

$$X_{n+1} := T_n^{n+1}.$$

Under the assumption on G, Lemma 14 applies which gives the relationship between $\partial_{\mathcal{W}_2} F_{n+1}$ and $\partial_{\mathcal{W}_2} G$. Together with the assumption on ξ_{n+1} by Assumption 1, we have that for each n, $\exists \eta_{n+1} \in \partial_{\mathcal{W}_2} G(p_{n+1})$ s.t.

$$\gamma \xi_{n+1} - \gamma \eta_{n+1} = \mathrm{I_d} - T_{n+1}^n, \quad p_{n+1}\text{-a.e.}$$

and equivalently,

$$\mathrm{I_d} - X_{n+1} = \gamma(\eta_{n+1} - \xi_{n+1}) \circ X_{n+1}. \quad p_n\text{-a.e.} \tag{4.52}$$

Denote by T_n^π the unique OT map from p_n to π. Expanding $\|X_{n+1} - T_n^\pi\|_{p_n}^2$ as

$$\begin{aligned}
\|X_{n+1} - T_n^\pi\|_{p_n}^2 &= \|(\mathrm{I_d} - X_{n+1}) - (\mathrm{I_d} - T_n^\pi)\|_{p_n}^2 \\
&= \|\mathrm{I_d} - T_n^\pi\|_{p_n}^2 - 2\langle \mathrm{I_d} - T_n^\pi, \mathrm{I_d} - X_{n+1}\rangle_{p_n} + \|\mathrm{I_d} - X_{n+1}\|_{p_n}^2 \\
&= \|\mathrm{I_d} - T_n^\pi\|_{p_n}^2 - 2\langle X_{n+1} - T_n^\pi, \mathrm{I_d} - X_{n+1}\rangle_{p_n} - \|\mathrm{I_d} - X_{n+1}\|_{p_n}^2 \\
&\leqslant \|\mathrm{I_d} - T_n^\pi\|_{p_n}^2 - 2\langle X_{n+1} - T_n^\pi, \mathrm{I_d} - X_{n+1}\rangle_{p_n},
\end{aligned}$$

where in the last inequality we use that $\|\mathrm{I_d} - X_{n+1}\|_{p_n}^2 \geqslant 0$. By that

$$\|\mathrm{I_d} - T_n^\pi\|_{p_n}^2 = \mathcal{W}_2(p_n, \pi)^2,$$

and together with (4.52), we have

$$\|X_{n+1} - T_n^\pi\|_{p_n}^2 \leqslant \mathcal{W}_2(p_n, \pi)^2 - 2\gamma\langle X_{n+1} - T_n^\pi, (\eta_{n+1} - \xi_{n+1}) \circ X_{n+1}\rangle_{p_n}. \tag{4.53}$$

Applying Lemma 6 with $p = p_n$ and $\rho = p_{n+1}$, we have

$$G(\pi) - G(p_{n+1}) \geqslant \langle T_n^\pi - X_{n+1}, \eta_{n+1} \circ X_{n+1}\rangle_{p_n} + \frac{\lambda}{2}\mathcal{W}_2(p_{n+1}, \pi)^2. \tag{4.54}$$

Meanwhile, by Cauchy Schwartz,

$$\begin{aligned}
|\langle X_{n+1} - T_n^\pi, \xi_{n+1} \circ X_{n+1}\rangle_{p_n}| &\leqslant \|X_{n+1} - T_n^\pi\|_{p_n} \|\xi_{n+1} \circ X_{n+1}\|_{p_n} \\
&\leqslant \varepsilon \|X_{n+1} - T_n^\pi\|_{p_n}
\end{aligned}$$

where the 2nd inequality is by that $\|\xi_{n+1} \circ X_{n+1}\|_{p_n} = \|\xi_{n+1}\|_{p_{n+1}} \leqslant \varepsilon$ (Assumption 1). Since $\lambda > 0$, we have

$$\varepsilon \|X_{n+1} - T_n^\pi\|_{p_n} \leqslant \frac{\varepsilon^2}{\lambda} + \frac{\lambda}{4}\|X_{n+1} - T_n^\pi\|_{p_n}^2. \tag{4.55}$$

Putting together, this gives

$$|\langle X_{n+1} - T_n^\pi, \xi_{n+1} \circ X_{n+1} \rangle_{p_n}| \leqslant \frac{\varepsilon^2}{\lambda} + \frac{\lambda}{4}\|X_{n+1} - T_n^\pi\|_{p_n}^2. \tag{4.56}$$

Inserting (4.54)(4.56) into (4.53) gives

$$(1 - \frac{\gamma\lambda}{2})\|X_{n+1} - T_n^\pi\|_{p_n}^2 \leqslant \mathcal{W}_2(p_n, \pi)^2 + 2\gamma \left(G(\pi) - G(p_{n+1}) - \frac{\lambda}{2}\mathcal{W}_2(p_{n+1}, \pi)^2 \right) + \frac{2\gamma}{\lambda}\varepsilon^2. \tag{4.57}$$

Because $(X_{n+1}, T_n^\pi)_\# p_n$ is a coupling between p_{n+1} and π, we have

$$\mathcal{W}_2(p_{n+1}, \pi)^2 \leqslant \|X_{n+1} - T_n^\pi\|_{p_n}^2 . \tag{4.58}$$

Under the condition of the lemma, $0 < \gamma\lambda < 2$, and thus $1 - \frac{\gamma\lambda}{2} > 0$ and then the l.h.s. of (4.57) $\geqslant (1 - \frac{\gamma\lambda}{2})\mathcal{W}_2(p_{n+1}, \pi)^2$. This proves (4.34). □

Proof of Theorem 8. Taking $\pi = q$ and apply Lemma 7, by that $2\gamma \left(G(p_{n+1}) - G(\pi)\right) \geqslant 0$, (4.34) gives that for all n,

$$\left(1 + \frac{\gamma\lambda}{2}\right) \mathcal{W}_2^2(p_{n+1}, q) \leqslant \mathcal{W}_2^2(p_n, q) + \frac{2\gamma}{\lambda}\varepsilon^2. \tag{4.59}$$

Define the numbers ρ and α as

$$\rho := \left(1 + \frac{\gamma\lambda}{2}\right)^{-1}, \quad 0 < \rho < 1, \quad \alpha := \sqrt{\frac{2\gamma}{\lambda}}\varepsilon,$$

and define

$$E_n := \mathcal{W}_2(p_n, q)^2,$$

then (4.59) can be written as

$$E_{n+1} \leqslant \rho(E_n + \alpha).$$

Recursively applying from 0 to $n-1$ gives that

$$E_n \leqslant \rho^n E_0 + \alpha \frac{\rho(1-\rho^n)}{1-\rho} \leqslant \rho^n E_0 + \alpha \frac{\rho}{1-\rho},$$

which by definition is equivalent to (4.35).

By (4.35), one will have $\mathcal{W}_2(p_n, q)^2 \leqslant 5\varepsilon^2/\lambda^2$ if

$$\left(1 + \frac{\gamma\lambda}{2}\right)^{-n} \mathcal{W}_2^2(p_0, q) \leqslant \frac{\varepsilon^2}{\lambda^2},$$

which is fulfilled as long as

$$n \geqslant \frac{2\left(\log \mathcal{W}_2(p_0, q) + \log(\lambda/\varepsilon)\right)}{\log(1 + \gamma\lambda/2)}.$$

This requirement of n is satisfied under (4.36) by that $0 < \gamma\lambda < 2$ and the elementary relation that $\log(1+x) \geqslant x/2$ for $x \in (0, 1)$. We have proved the $\mathcal{W}_2$-error bound.

To show the smallness of the objective gap $G(p_n) - G(q)$, we use (4.34) again, and by that $\mathcal{W}_2^2(p_{n+1}, q) \geqslant 0$,

$$2\gamma\left(G(p_{n+1}) - G(q)\right) \leqslant \mathcal{W}_2^2(p_n, q) + \frac{2\gamma}{\lambda}\varepsilon^2. \tag{4.60}$$

When n already makes $\mathcal{W}_2(p_n, q)^2 \leqslant 5\varepsilon^2/\lambda^2$, we have

$$2\gamma\left(G(p_{n+1}) - G(q)\right) \leqslant (5 + 2\gamma\lambda)\frac{\varepsilon^2}{\lambda^2} \leqslant 9\frac{\varepsilon^2}{\lambda^2}, \tag{4.61}$$

where in the 2nd inequality we use that $\gamma\lambda < 2$ because $0 < \lambda \leqslant 1$ and $0 < \gamma < 2$. This proves the bound of $G(p_n) - G(q)$ in (4.37). □

4.8 Proofs in Section 4.5

4.8.1 Proofs in Section 4.5.1

Proof of Lemma 9. Let $X_1 \sim p$, $X_2 \sim q$, and

$$Y_1 = T(X_1), \quad Y_2 = T(X_2).$$

Then Y_1 and Y_2 also have densities, $Y_1 \sim \tilde{p} := T_\# p$ and $Y_2 \sim \tilde{q} := T_\# q$. By the data processing inequality concerning two probability distributions through the same stochastic transformation for the KL divergence (see, e.g., the introduction of Raginsky, 2016),

$$\mathrm{KL}(\tilde{p}||\tilde{q}) \leqslant \mathrm{KL}(p||q).$$

In the other direction, $X_i = T^{-1}(Y_i)$, $i = 1, 2$, then data processing inequality also implies

$$\mathrm{KL}(p||q) \leqslant \mathrm{KL}(\tilde{p}||\tilde{q}).$$

□

Proof of Corollary 10. Under Assumption 4, $q \in \mathcal{P}_2^r$, and then $G(q) = 0$ is the global minimum of G. Apply Theorem 8, for the N defined in the corollary, we have

$$G(p_N) \leqslant \frac{9}{2\gamma} \left(\frac{\varepsilon}{\lambda}\right)^2.$$

By Assumption 1, T_1^N as defined in (4.38) is invertible and has a finite Lipschitz constant on $\mathbb{R}^d$, and so is its inverse $(T_1^N)^{-1}$. In the forward direction, since $p_0 \in \mathcal{P}_2^r$, $p_N = (T_1^N)_\# p_0$ is in $\mathcal{P}_2^r$ by Lemma 5. Similarly, since $q_N = q \in \mathcal{P}_2^r$, $q_0 = ((T_1^N)^{-1})_\# q_N$ is also in $\mathcal{P}_2^r$. Then Lemma 9 gives that $\mathrm{KL}(p_0||q_0) = \mathrm{KL}(p_N||q_N) = \mathrm{KL}(p_N||q) = G(p_N)$ which is bounded as stated in the corollary. The TV bound follows by the Pinsker's inequality. □

4.8.2 Proofs in Section 4.5.1

Lemma 18 (ρ_δ and $\mathcal{W}_2$ closeness)**.** *Suppose $P \in \mathcal{P}_2$, and ρ_t is the density of X_t in an OU process as in* (4.6)*, then*

(i) $\rho_t \in \mathcal{P}_2^r$ for any $t > 0$,

(ii) $\forall \varepsilon > 0$, $\exists \delta > 0$ s.t. $\mathcal{W}_2(\rho_\delta, P) < \varepsilon$. In this case, one can choose $\delta \sim \varepsilon^2$.

Proof of Lemma 18. For the OU process, we have $V(x) = \|x\|^2/2$ in (4.7). Then for any $t > 0$, $\rho_t = \mathcal{L}_t(P)$ has the expression as

$$\rho_t(x) = \int_{\mathbb{R}^d} \frac{1}{(2\pi\sigma_t^2)^{d/2}} e^{-\|x - e^{-t}y\|^2/(2\sigma_t^2)} dP(y), \quad \sigma_t^2 := 1 - e^{-2t}. \tag{4.62}$$

Equivalently, ρ_t is the probability density of the random vector

$$Z_t := e^{-t}X_0 + \sigma_t Z, \quad Z \sim \mathcal{N}(0, I_d), \quad Z \text{ is independent from } X_0.$$

Since $\mathbb{E}\|Z_t\|^2 = e^{-2t}M_2(P) + \sigma_t^2 d < \infty$, we have $\rho_t \in \mathcal{P}_2$ and this proves (i).

To prove (ii): Because the law of (Z_t, X_0) is a coupling of ρ_t and P,

$$\begin{aligned}
\mathcal{W}_2(\rho_t, P)^2 &\leqslant \mathbb{E}\|Z_t - X_0\|^2 \\
&= \mathbb{E}\|(e^{-t} - 1)X_0 + \sigma_t Z\|^2 \\
&= (1 - e^{-t})^2 M_2(P) + (1 - e^{-2t})d \\
&\leqslant t^2 M_2(P) + 2td,
\end{aligned}$$

where in the last inequality we used that $1 - e^{-x} \leqslant x, \forall x \geqslant 0$. Since $M_2(P) < \infty$, we have bounded $\mathcal{W}_2(\rho_t, P)^2$ to be $O(t)$. □

Proof of Corollary 11. For the ε in Assumption 1, the existence of δ to make $\mathcal{W}_2(\rho_\delta, P) < \varepsilon$ is by Lemma 18, and we also have $\rho_\delta \in \mathcal{P}_2^r$. The rest of the proof is the same as in Corollary 10 by starting from $p_0 = \rho_\delta$. □

4.8.3 Proofs and lemmas in Section 4.5.2

Lemma 19 (Lipschitz bound of ODE solution map)**.** *Suppose for $\gamma > 0$, $\hat{v}(x, t)$ is C^1 in (x, t) and Lipschitz in x uniformly on $\mathbb{R}^d \times [0, \gamma]$ with Lipschitz constant $K \geqslant 0$. Let $x(t)$ be the solution to the ODE*

$$\dot{x}(t) = \hat{v}(x(t), t), \quad t \in [0, \gamma], \tag{4.63}$$

and define the solution map from 0 to γ as $T : \mathbb{R}^d \to \mathbb{R}^d$, that is,

$$T(x_0) = x_0 + \int_0^\gamma \hat{v}(x(t), t)dt, \quad x(0) = x_0. \tag{4.64}$$

Then T is invertible on $\mathbb{R}^d$, and both T and T^{-1} are Lipschitz on $\mathbb{R}^d$ with Lipschitz constant $e^{\gamma K}$.

Proof of Lemma 19. Let $x_1(t)$ and $x_2(t)$ be the solution to the ODE (4.63) from $x_1(0) = y$, and $x_2(0) = z$ respectively. By definition,

$$T(y) = x_1(\gamma), \quad T(z) = x_2(\gamma).$$

Under the condition of $\hat{v}$, the ODE is well-posed Sideris, 2013. This implies the invertibility of T, and T^{-1} is the solution map of the reverse time ODE from $t = \gamma$ to $t = 0$.

We now prove the Lipschitz constant of T on $\mathbb{R}^d$, and that of T^{-1} can be proved similarly by considering the reverse time ODE. We want to show that

$$\|T(y) - T(z)\| \leqslant e^{\gamma K}\|y - z\|, \quad \forall y, z \in \mathbb{R}^d,$$

and this is equivalent to that for any $x_1(0),\ x_2(0) \in \mathbb{R}^d$,

$$\|x_1(\gamma) - x_2(\gamma)\| \leqslant e^{\gamma K}\|x_1(0) - x_2(0)\|. \tag{4.65}$$

For fixed $x_1(0),\ x_2(0)$, define

$$E(t) := \frac{1}{2}\|x_1(t) - x_2(t)\|^2,$$

then $E(0) = \|x_1(0) - x_2(0)\|^2/2$, and

$$\begin{aligned}\dot{E}(t) &= (x_1(t) - x_2(t))^T(\dot{x}_1(t) - \dot{x}_2(t)) \\ &= (x_1(t) - x_2(t))^T(\hat{v}(x_1(t), t) - \hat{v}(x_2(t), t)).\end{aligned}$$

Thus, by that $\|\hat{v}(x_1(t), t) - \hat{v}(x_2(t), t)\| \leqslant K\|x_1(t) - x_2(t)\|$, we have

$$\dot{E}(t) \leqslant K\|x_1(t) - x_2(t)\|^2 = 2KE(t).$$

By Grönwall, $E(t) \leqslant E(0)e^{2Kt}$, and this gives

$$\|x_1(t) - x_2(t)\|^2 \leqslant e^{2Kt}\|x_1(0) - x_2(0)\|^2, \quad t \in [0, \gamma].$$

Setting $t = \gamma$ proves (4.65). □

Proof of Proposition 12. By construction, for $n = 1, \cdots, N$,

$$q_{n-1} = (T_n^{-1})_{\#}q_n, \quad \tilde{q}_{n-1} = (S_n)_{\#}\tilde{q}_n.$$

Since $q_N = q$, $q \in \mathcal{P}_2^r$ by Assumption 4, and for any n,

$$q_{n-1} = (T_n^{-1} \circ \cdots \circ T_N^{-1})_{\#}q, \quad \tilde{q}_{n-1} = (S_n \circ \cdots \circ S_N)_{\#}q,$$

where $T_n^{-1} \circ \cdots \circ T_N^{-1}$ $(S_n \circ \cdots \circ S_N)$ is globally Lipschitz on $\mathbb{R}^d$ (by Assumption 6), then $q_{n-1} \in \mathcal{P}_2^r$ $(\tilde{q}_{n-1} \in \mathcal{P}_2^r)$ by Lemma 5. Thus we have that q_n and $\tilde{q}_n$ are all in $\mathcal{P}_2^r$.

For each n, we have

$$\begin{aligned}\mathcal{W}_2(\tilde{q}_{n-1}, q_{n-1}) &= \mathcal{W}_2((S_n)_\# \tilde{q}_n, (T_n^{-1})_\# q_n) \\ &\leqslant \mathcal{W}_2((S_n)_\# \tilde{q}_n, (T_n^{-1})_\# \tilde{q}_n) + \mathcal{W}_2((T_n^{-1})_\# \tilde{q}_n, (T_n^{-1})_\# q_n) \\ &:= ① + ②.\end{aligned}$$

To bound ①, we use Assumption 2. Define $L := e^{\gamma K}$. Using $(S_n, T_n^{-1})_\# \tilde{q}_n$ as the coupling, we have

$$\begin{aligned}\mathcal{W}_2^2((S_n)_\# \tilde{q}_n, (T_n^{-1})_\# \tilde{q}_n) &\leqslant \int_{\mathbb{R}^d} \|S_n(x) - T_n^{-1}(x)\|^2 \tilde{q}_n(x) dx \\ &\leqslant \int_{\mathbb{R}^d} L^2 \|T_n \circ S_n(x) - x\|^2 \tilde{q}_n(x) dx \quad (T_n^{-1} \text{ is } L\text{-Lipschitz}) \\ &= L^2 \|T_n \circ S_n - \mathrm{I_d}\|_{\tilde{q}_n}^2,\end{aligned}$$

and thus

$$① \leqslant L\varepsilon_{\text{inv}}. \tag{4.66}$$

To bound ②, we use that T_n^{-1} is L-Lipschitz on $\mathbb{R}^d$ again. Specifically, let Y_n be the unique OT map from q_n to $\tilde{q}_n$ which is well-defined by the Brenier Theorem, then $(T_n^{-1} \circ Y_n, T_n^{-1})_\# q_n$ is a coupling of $(T_n^{-1})_\# \tilde{q}_n$ and $(T_n^{-1})_\# q_n$. We have that

$$\begin{aligned}\mathcal{W}_2^2((T_n^{-1})_\# \tilde{q}_n, (T_n^{-1})_\# q_n) &\leqslant \int_{\mathbb{R}^d} \|T_n^{-1} \circ Y_n(x) - T_n^{-1}(x)\|^2 q_n(x) dx \\ &\leqslant \int_{\mathbb{R}^d} L^2 \|Y_n(x) - x\|^2 q_n(x) dx \\ &= L^2 \mathcal{W}_2^2(\tilde{q}_n, q_n).\end{aligned}$$

Thus

$$② \leqslant L\mathcal{W}_2(\tilde{q}_n, q_n). \tag{4.67}$$

Putting together, we have

$$\mathcal{W}_2(\tilde{q}_{n-1}, q_{n-1}) \leqslant e^{\gamma K}(\varepsilon_{\text{inv}} + \mathcal{W}_2(\tilde{q}_n, q_n)).$$

Note that $\mathcal{W}_2(\tilde{q}_N, q_N) = 0$ by that $\tilde{q}_N = q_N = q$. Applying recursively from $n = N$ to $n = 1$ gives that

$$\mathcal{W}_2(\tilde{q}_0, q_0) \leqslant \varepsilon_{\text{inv}} \frac{e^{\gamma K}(e^{\gamma K N} - 1)}{e^{\gamma K} - 1}, \tag{4.68}$$

which proves (4.41) by that $e^x - 1 \geqslant x$ for $x \geqslant 0$. □

Proof of Corollary 13. Under the condition of the corollary, Corollary 10 applies to bound $\mathrm{KL}(p||q_0)$ and $TV(p, q_0)$ as in (4.40), and Proposition 12 applies to bound $\mathcal{W}_2(\tilde{q}_0, q_0)$ as in (4.41). It suffices to show that the r.h.s. of (4.41) is less than or equal to that of (4.42).

By the choice of N in (4.39),

$$N \leqslant \frac{8}{\gamma\lambda} \left(\log \mathcal{W}_2(p_0, q) + \log(\lambda/\varepsilon)\right) + 1,$$

and thus

$$e^{\gamma K(N+1)} \leqslant e^{2\gamma K} \left(\mathcal{W}_2(p_0, q) \frac{\lambda}{\varepsilon}\right)^{8K/\lambda},$$

which proves the needed inequality. □

5. Conclusions

In conclusion, the thesis made significant contributions to the field of generative modeling by analyzing the convergence properties of diffusion-based sampling methods and related flow-based models. By focusing on diffusion models and their extensions, we have provided novel insights into the theoretical foundations of these methods, paving the way for their broader application and understanding.

For the diffusion models, we present several interesting further directions to explore. First, the assumption that we have a score estimate that is $O(1)$-accurate in L^2, although weaker than the usual assumptions for theoretical analysis, is in fact still a strong condition in practice that seems unlikely to be satisfied (and difficult to check) when learning complex distributions such as distributions of images. What would a reasonable weaker condition be, and in what sense can we still obtain reasonable samples? Second, our analysis assumes a L^2-estimate of the score function is given, but the question remains of when we can find such an estimate. What natural conditions on distributions allow their score functions to be learned by a neural network? Various works have considered the representability of data distributions by diffusion-like processes (Tzen & Raginsky, 2019), but the questions of optimization and generalization appear more challenging.

The work about flow-based method can be extended in several directions. First, the assumption on the learning in the forward process, Assumption 1, is theoretical and cannot be directly verified in practice. Any relaxation of the assumption would be interesting and may provide guidance on training such flow networks. Second, the current generation result only covers the case of G being the KL divergence. An extension to the cases when G is other types of divergence, such as f-divergence, will broaden the scope of the result. Third, our theory uses the population quantities throughout. A finite-sample analysis, which can be based on our population analysis, will provide statistical convergence rates in addition to the current result.

Meanwhile, the JKO scheme computes a fully backward proximal GD. Given the existing convergence rates of the other Wasserstein GD (Kent et al., 2021; Salim et al., 2020),

one would expect that a variety of first-order Wasserstein GD schemes can be applied to progressive flow models and the theoretical guarantees can be derived similarly to the JKO scheme. We also note the connection between the JKO scheme and learning of the score function, at least in the limit of small step size (Xu et al., 2023b, Section 3.2). Given the growing literature on the analysis of score-based diffusion models, it can be worthwhile to further investigate this connection to develop new theories for the ODE flow models.

Bibliography

Albergo, M. S., Boffi, N. M., & Vanden-Eijnden, E. (2023). Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*.

Albergo, M. S., & Vanden-Eijnden, E. (2023). Building normalizing flows with stochastic interpolants. *The Eleventh International Conference on Learning Representations*.

Alvarez-Melis, D., Schiff, Y., & Mroueh, Y. (2021). Optimizing functionals on the space of probabilities with input convex neural networks. *arXiv preprint arXiv:2106.00774*.

Amari, S.-i. (2008). Information geometry and its applications: Convex function and dually flat manifold. *LIX Fall Colloquium on Emerging Trends in Visual Computing*, 75–102.

Amari, S.-i. (2016). *Information geometry and its applications* (Vol. 194). Springer.

Ambrosio, L., Gigli, N., & Savaré, G. (2005). *Gradient flows: In metric spaces and in the space of probability measures*. Springer Science & Business Media.

Amos, B., Xu, L., & Kolter, J. Z. (2017). Input convex neural networks. *International Conference on Machine Learning*, 146–155.

Anderson, B. D. (1982). Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, *12*(3), 313–326.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. *International conference on machine learning*, 214–223.

Arora, S., Risteski, A., & Zhang, Y. (2018). Do gans learn the distribution? some theory and empirics. *International Conference on Learning Representations*.

Bakry, D., Gentil, I., & Ledoux, M. (2013). *Analysis and geometry of Markov diffusion operators* (Vol. 348). Springer Science & Business Media.

Behrmann, J., Grathwohl, W., Chen, R. T., Duvenaud, D., & Jacobsen, J.-H. (2019). Invertible residual networks. *International Conference on Machine Learning*, 573–582.

Benton, J., De Bortoli, V., Doucet, A., & Deligiannidis, G. (2023). Linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*.

Benton, J., Deligiannidis, G., & Doucet, A. (2023). Error bounds for flow matching methods. *arXiv preprint arXiv:2305.16860*.

Bernton, E. (2018). Langevin monte carlo and JKO splitting. *Proceedings of the 31st Conference On Learning Theory*, *75*, 1777–1798.

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

Bolley, F., Gentil, I., & Guillin, A. (2012). Convergence to equilibrium in wasserstein distance for fokker–planck equations. *Journal of Functional Analysis*, *263*(8), 2430–2457.

Brascamp, H. J., & Lieb, E. H. (2002). On extensions of the brunn-minkowski and prékopa-leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. In *Inequalities* (pp. 441–464). Springer.

Chafaï, D. (2004). Entropies, convexity, and functional inequalities, on Φ-entropies and Φ-sobolev inequalities. *Journal of Mathematics of Kyoto University*, *44*(2), 325–363.

Chen, H.-B., Chewi, S., & Niles-Weed, J. (2021). Dimension-free log-sobolev inequalities for mixture distributions. *Journal of Functional Analysis*, *281*(11), 109236.

Chen, H., Lee, H., & Lu, J. (2023). Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. *International Conference on Machine Learning*, 4735–4763.

Chen, R. T., Rubanova, Y., Bettencourt, J., & Duvenaud, D. K. (2018). Neural ordinary differential equations. *Advances in neural information processing systems*, *31*.

Chen, S., Chewi, S., Lee, H., Li, Y., Lu, J., & Salim, A. (2023). The probability flow ode is provably fast. *arXiv preprint arXiv:2305.11798*.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., & Zhang, A. R. (2022). Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions [arXiv:2209.11215].

Chen, S., Daras, G., & Dimakis, A. (2023). Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for ddim-type samplers. *International Conference on Machine Learning*, 4462–4484.

Cheng, X., & Bartlett, P. (2018). Convergence of langevin mcmc in kl-divergence. *Algorithmic Learning Theory*, 186–211.

Cheng, X., Chatterji, N. S., Bartlett, P. L., & Jordan, M. I. (2018). Underdamped langevin mcmc: A non-asymptotic analysis. *Conference on learning theory*, 300–323.

Cheng, X., & Cloninger, A. (2022). Classification logit two-sample testing by neural networks for differentiating near manifold densities. *IEEE Transactions on Information Theory*, *68*(10), 6631–6662.

Cheng, X., Lu, J., Tan, Y., & Xie, Y. (2023). Convergence of flow-based generative models via proximal gradient descent in wasserstein space.

Chewi, S., Erdogdu, M. A., Li, M. B., Shen, R., & Zhang, M. (2021). Analysis of langevin monte carlo from poincaré to log-sobolev. *arXiv preprint arXiv:2112.12662*.

Dalalyan, A. S. (2017). Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*(3), 651–676.

Dalalyan, A. S., & Karagulyan, A. (2019). User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *Stochastic Processes and their Applications*, *129*(12), 5278–5311.

Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., Yosinski, J., & Liu, R. (2019). Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

De Bortoli, V. (2022). Convergence of denoising diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2208.05314*.

De Bortoli, V., Thornton, J., Heng, J., & Doucet, A. (2021). Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, *34*.

Degond, P., & Mas-Gallic, S. (1989). The weighted particle method for convection-diffusion equations. i. the case of an isotropic viscosity. *Mathematics of computation*, *53*(188), 485–507.

Degond, P., & Mustieles, F.-J. (1990). A deterministic approximation of diffusion equations using particles. *SIAM Journal on Scientific and Statistical Computing*, *11*(2), 293–310.

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, *34*.

Diao, M. Z., Balasubramanian, K., Chewi, S., & Salim, A. (2023). Forward-backward gaussian variational inference via JKO in the Bures-Wasserstein space. *International Conference on Machine Learning*, 7960–7991.

Dinh, L., Krueger, D., & Bengio, Y. (2014). NICE: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*.

Dinh, L., Sohl-Dickstein, J., & Bengio, S. (2016). Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*.

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

Durmus, A., & Moulines, E. (2017). Nonasymptotic convergence analysis for the unadjusted langevin algorithm. *The Annals of Applied Probability*, *27*(3), 1551–1587.

Erdogdu, M. A., Hosseinzadeh, R., & Zhang, M. S. (2021). Convergence of langevin monte carlo in chi-squared and renyi divergence. *arXiv preprint arXiv:2007.11612*.

Fan, J., Zhang, Q., Taghvaei, A., & Chen, Y. (2022). Variational wasserstein gradient flow. *International Conference on Machine Learning*, 6185–6215.

Finlay, C., Jacobsen, J.-H., Nurbekyan, L., & Oberman, A. (2020). How to train your neural ODE: The world of jacobian and kinetic regularization. *International conference on machine learning*, 3154–3164.

Ge, R., Lee, H., & Risteski, A. (2018). Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering langevin monte carlo. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 7858–7867.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

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

Grathwohl, W., Chen, R. T., Bettencourt, J., Sutskever, I., & Duvenaud, D. (2018). Ffjord: Free-form continuous dynamics for scalable reversible generative models. *International Conference on Learning Representations*.

Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D., Norouzi, M., & Swersky, K. (2019). Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of wasserstein gans. *NIPS*.

Hargé, G. (2004). A convex/log-concave correlation inequality for gaussian measure and an application to abstract wiener spaces. *Probability theory and related fields*, *130*(3), 415–440.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, *33*, 6840–6851.

Huang, H., Yu, J., Chen, J., & Lai, R. (2023). Bridging mean-field games and normalizing flows with trajectory regularization. *Journal of Computational Physics*, 112155.

Huang, J., Jiao, Y., Li, Z., Liu, S., Wang, Y., & Yang, Y. (2022). An error analysis of generative adversarial networks for learning distributions. *The Journal of Machine Learning Research*, *23*(1), 5047–5089.

Hyvärinen, A., & Dayan, P. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, *6*(4).

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5967–5976.

Jing, B., Corso, G., Berlinghieri, R., & Jaakkola, T. (2022). Subspace diffusion generative models. *arXiv preprint arXiv:2205.01490*.

Johnson, R., & Zhang, T. (2019). A framework of composite functional gradient methods for generative adversarial models. *IEEE transactions on pattern analysis and machine intelligence*, *43*(1), 17–32.

Jordan, R., Kinderlehrer, D., & Otto, F. (1998). The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, *29*(1), 1–17.

Kent, C., Li, J., Blanchet, J., & Glynn, P. W. (2021). Modified frank wolfe in probability space. *Advances in Neural Information Processing Systems*, *34*, 14448–14462.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

Kingma, D. P., & Welling, M. (2019a). An introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning*, *12*(4), 307–392. https://doi.org/10.1561/2200000056

Kingma, D. P., & Welling, M. (2019b). An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*.

Kingma, D. P., & Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, *31*.

Kobyzev, I., Prince, S. J., & Brubaker, M. A. (2020). Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, *43*(11), 3964–3979.

Lambert, M., Chewi, S., Bach, F., Bonnabel, S., & Rigollet, P. (2022). Variational inference via Wasserstein gradient flows. *Advances in Neural Information Processing Systems*, *35*, 14434–14447.

Laurent, B., & Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 1302–1338.

Lee, H., Ge, R., Ma, T., Risteski, A., & Arora, S. (2017). On the ability of neural nets to express distributions. *Conference on Learning Theory*, 1271–1296.

Lee, H., Lu, J., & Tan, Y. (2022). Convergence for score-based generative modeling with polynomial complexity. *arXiv preprint arXiv:2206.06227*.

Lee, H., Lu, J., & Tan, Y. (2023). Convergence of score-based generative modeling for general data distributions. *International Conference on Algorithmic Learning Theory*, 946–985.

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

Li, G., Wei, Y., Chen, Y., & Chi, Y. (2023). Towards faster non-asymptotic convergence for diffusion-based generative models. *arXiv preprint arXiv:2306.09251*.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., & Le, M. (2023). Flow matching for generative modeling. *The Eleventh International Conference on Learning Representations*.

Liu, Q. (2022). Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*.

Lu, Y., & Lu, J. (2020). A universal approximation theorem of deep neural networks for expressing probability distributions. *Advances in neural information processing systems*, *33*, 3094–3105.

Majka, M. B., Mijatović, A., & Szpruch, Ł. (2020). Nonasymptotic bounds for sampling algorithms without log-concavity. *The Annals of Applied Probability*, *30*(4), 1534–1581.

Marzouk, Y., Ren, Z., Wang, S., & Zech, J. (2023). Distribution learning via neural differential equations: A nonparametric statistical perspective. *arXiv preprint arXiv:2309.01043*.

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., & Ermon, S. (2021). Sdedit: Guided image synthesis and editing with stochastic differential equations. *International Conference on Learning Representations*.

Mokrov, P., Korotin, A., Li, L., Genevay, A., Solomon, J. M., & Burnaev, E. (2021). Large-scale wasserstein gradient flows. *Advances in Neural Information Processing Systems*, *34*, 15243–15256.

Nemirovski, A., Juditsky, A., Lan, G., & Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, *19*(4), 1574–1609.

Nemirovsky, A., & Yudin, D. (1983). Problem complexity and method efficiency in optimization.

Onken, D., Wu Fung, S., Li, X., & Ruthotto, L. (2021). OT-flow: Fast and accurate continuous normalizing flows via optimal transport. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(1-18).

Pedrotti, F., Maas, J., & Mondelli, M. (2023). Improved convergence of score-based diffusion models via prediction-correction. *arXiv preprint arXiv:2305.14164*.

Perekrestenko, D., Eberhard, L., & Bölcskei, H. (2021). High-dimensional distribution generation through deep neural networks. *Partial Differential Equations and Applications*, *2*(5), 1–44.

Raginsky, M. (2016). Strong data processing inequalities and Φ-sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, *62*(6), 3355–3389.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Ruthotto, L., Osher, S. J., Li, W., Nurbekyan, L., & Fung, S. W. (2020). A machine learning framework for solving high-dimensional mean field game and mean field control problems. *Proceedings of the National Academy of Sciences*, *117*(17), 9183–9193.

Salim, A., Korba, A., & Luise, G. (2020). The wasserstein proximal gradient algorithm. *Advances in Neural Information Processing Systems*, *33*, 12356–12366.

Sideris, T. C. (2013). Ordinary differential equations and dynamical systems.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. *International Conference on Machine Learning*, 2256–2265.

Song, Y., Durkan, C., Murray, I., & Ermon, S. (2021). Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, *34*.

Song, Y., & Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems*.

Song, Y., & Ermon, S. (2020). Improved techniques for training score-based generative models. *arXiv preprint arXiv:2006.09011*.

Song, Y., Garg, S., Shi, J., & Ermon, S. (2020). Sliced score matching: A scalable approach to density and score estimation. *Uncertainty in Artificial Intelligence*, 574–584.

Song, Y., Shen, L., Xing, L., & Ermon, S. (2021). Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*.

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

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021). Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*.

Sreekumar, S., & Goldfeld, Z. (2022). Neural estimation of statistical divergences. *Journal of machine learning research*, *23*(126).

Tzen, B., & Raginsky, M. (2019). Theoretical guarantees for sampling and inference in generative models with latent diffusions. *Conference on Learning Theory*, 3084–3114.

Vempala, S., & Wibisono, A. (2019). Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, *32*, 8094–8106.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science* (Vol. 47). Cambridge university press.

Vidal, A., Wu Fung, S., Tenorio, L., Osher, S., & Nurbekyan, L. (2023). Taming hyperparameter tuning in continuous normalizing flows using the jko scheme. *Scientific Reports*, *13*(1), 4501.

Vincent, P. (2011). A connection between score matching and denoising autoencoders. *Neural computation*, *23*(7), 1661–1674.

Xu, C., Cheng, X., & Xie, Y. (2022). Invertible neural networks for graph prediction. *IEEE Journal on Selected Areas in Information Theory*, *3*(3), 454–467. https://doi.org/10.1109/JSAIT.2022.3221864

Xu, C., Cheng, X., & Xie, Y. (2023a). Computing high-dimensional optimal transport by flow neural networks. *arXiv preprint arXiv:2305.11857*.

Xu, C., Cheng, X., & Xie, Y. (2023b). Normalizing flow neural networks by JKO scheme. *Conference on Neural Information Processing Systems (NeurIPS)*.

Yang, Y., Li, Z., & Wang, Y. (2022). On the capacity of deep generative networks for approximating distributions. *Neural networks*, *145*, 144–154.

Zhang, Q., & Chen, Y. (2022). Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*.

Zhao, J., Mathieu, M., & LeCun, Y. (2016). Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*.

Biography

Yixin Tan is a PhD candidate in Mathematics at Duke University, where he has pursued research driven by a passion for exploring the intersection of advanced mathematical concepts and cutting-edge applications in machine learning. He received his Bachelor's degree in Mathematics from Wuhan University.

Throughout his academic journey, Yixin has demonstrated a keen interest in theoretical and applied aspects of mathematics, with a particular focus on generative modeling and diffusion-based sampling methods. His research has contributed to advancing the understanding of these methods, culminating in several publications in top-tier conferences.

However, Yixin is not good at writing the biography.