BROADENING TARGET DISTRIBUTIONS FOR ACCELERATED DIFFUSION MODELS VIA A NOVEL ANALYSIS APPROACH

Anonymous authors
Paper under double-blind review

ABSTRACT

Accelerated diffusion models hold the potential to significantly enhance the efficiency of standard diffusion processes. Theoretically, these models have been shown to achieve faster convergence rates than the standard $\mathcal{O}(1/\epsilon^2)$ rate of vanilla diffusion models, where ϵ denotes the target accuracy. However, current theoretical studies have established the acceleration advantage only for restrictive target distribution classes, such as those with smoothness conditions imposed along the entire sampling path or with bounded support. In this work, we significantly broaden the target distribution classes with a new accelerated stochastic DDPM sampler. In particular, we show that it achieves accelerated performance for three broad distribution classes not considered before. Our first class relies on the smoothness condition posed only to the target density q_0, which is far more relaxed than the existing smoothness conditions posed to all q_t along the entire sampling path. Our second class requires only a finite second moment condition, allowing for a much wider class of target distributions than the existing finite-support condition. Our third class is Gaussian mixture, for which our result establishes the first acceleration guarantee. Moreover, among accelerated DDPM type samplers, our results specialized for bounded-support distributions show an improved dependency on the data dimension d. Our analysis introduces a novel technique for establishing performance guarantees via constructing a tilting factor representation of the convergence error and utilizing Tweedie's formula to handle Taylor expansion terms. This new analytical framework may be of independent interest.

1 INTRODUCTION

Generative modeling is a fundamental task in machine learning, aiming to generate samples out of a distribution similar to that of training data. Classical generative models include variational autoencoders (VAE) (Kingma & Welling, 2022), generative adversarial networks (GANs) (Goodfellow et al., 2014), and normalizing flows Rezende & Mohamed (2015), etc. Recently, diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song & Ermon, 2019) have arisen as an appealing generative model and have received wide popularity due to their excellent performance over a variety of tasks and applications as summarized in many surveys of diffusion models (Yang et al., 2023; Croitoru et al., 2023; Kazerouni et al., 2023).

The empirical success of diffusion models has also inspired extensive theoretical studies, aiming to characterize the convergence guarantee for diffusion models. The convergence rate (i.e., the total number of steps to attain a target accuracy ε) for standard vanilla Denoising Diffusion Probabilistic Models (DDPMs) has been established to be $\mathcal{O}(\varepsilon^{-2})$ for wide classes of target distributions (Chen et al., 2023a; Benton et al., 2024a; Conforti et al., 2023) (see Appendix A for a more complete summary). More recently, various **accelerated** samplers have been proposed and been shown to achieve an improved convergence rate of $\mathcal{O}(\varepsilon^{-1})$. One such acceleration approach is to redesign the (stochastic) DDPM reverse process. This includes augmenting the original reverse process with an additional estimate (Li et al., 2024c), introducing intermediate sampling points along the generation path (Li et al., 2024a), and employing special Markov-chain Monte-Carlo (MCMC) algorithms (Huang et al., 2024b). Another acceleration method is to sample with the corresponding probability ODE (Li et al., 2024c; Chen et al., 2023c; Huang et al., 2024a; Li et al., 2024d).

Target distribution Q_0	**Method**	**Num of steps**	**Results**
$\nabla \log q_t, s_t$ L-Lips. $\forall t$	ODE-based	$\mathcal{O}\left(\frac{\sqrt{d}L^2}{\varepsilon}\right)$	(Chen et al., 2023c, Thm 3)
$\nabla \log q_t$ L-Lips. $\forall t$	DDPM accl.	$\mathcal{O}\left(\frac{\sqrt{d}L^2}{\varepsilon}\right)$	(Huang et al., 2024b, Thm 4.4)†
$\left\|\partial_{\boldsymbol{a}}^k s_t(x)\right\| \leq L\ \forall x, t, \boldsymbol{a}$ and $\forall k \leq p+1$, Q_0 Bounded Support	ODE	$\mathcal{O}\left(\frac{d^{\frac{p+1}{p}}}{\varepsilon^{\frac{1}{p}}}\right)^*$	(Huang et al., 2024a, Thm 3.10)†
$\nabla^2 \log q_0$ M-Lips.	DDPM accl.	$\mathcal{O}\left(\frac{d^{1.5}\log^{1.5} M}{\varepsilon}\right)$	(This paper, Thm 4)
Q_0 Gaussian Mixture	DDPM accl.	$\mathcal{O}\left(\frac{d^{1.5}N^{1.5}}{\varepsilon}\right)$	(This paper, Thm 2)
Q_0 Bounded Support	DDPM accl.	$\mathcal{O}\left(\frac{d^3}{\varepsilon}\right)^*$	(Li et al., 2024c, Thm 4) (Li et al., 2024a, Thm 2)†
	ODE	$\mathcal{O}\left(\frac{d^3}{\sqrt{\varepsilon}}\right)^*$	(Li et al., 2024c, Thm 2) (Li et al., 2024a, Thm 1)†
	ODE	$\mathcal{O}\left(\frac{d^2}{\varepsilon}\right)^*$	(Li et al., 2024c, Thm 1)
Q_0 Finite Variance	DDPM accl.	$\mathcal{O}\left(\frac{d^{1.5}}{\varepsilon}\right)^*$	(This paper, Thm 3)

Table 1: *Summary of accelerated convergence results in terms of the number of steps needed to achieve ε-accuracy in total variation, where d is the dimension. For Gaussian mixture, assume that $N \leq d$. The first 4 rows of this table correspond to the results under those target distributions with some smoothness conditions imposed, while the last 4 rows correspond to the results under (possibly) non-smooth targets with finite variance.* $(*)$ *Those results correspond to an* early-stopped *procedure that compares the sampling distribution to $Q_1(\delta)$, where $\mathrm{W}_2\left(Q_0, Q_1\right)^2 \lesssim \delta d$. Here the dependencies on δ are omitted.* $(\dagger)$ *Those studies are concurrent to our work based on the time that they were posted on arXiv. Note that this table does not include the studies within two months of the conference submission, but those are discussed in the related works.*

However, existing results on the acceleration guarantee suffer from strong assumptions on the target distribution. (i) For smooth target distributions, the analyses of Chen et al. (2023c); Huang et al. (2024a;b) require that all the scores (or their close estimates or both) satisfy certain Lipschitz-smooth condition *along the entire sampling path*, i.e., the smoothness condition is posed to the density q_t for all iteration time t. However, such smoothness at intermediate steps is generally restrictive and hard to verify in practice. (ii) For (possibly) non-smooth targets, the analysis of Li et al. (2024a;c;d) requires the distribution to have finite support for early-stopped sampling procedures. Such an assumption is, however, restrictive if compared to that for early-stopped vanilla samplers, where convergence guarantees have been established only under the assumption of finite variance (Chen et al., 2023a; Benton et al., 2024a). The above discussions raise the following important open question:

Question 1: Can we obtain an accelerated convergence rate for a much broader set of target distributions? Namely, for smooth target distributions, can the smoothness condition be imposed only on the target distribution; and for (possibly) non-smooth targets, can we broaden the target distribution only to have finite variance?

Further, the existing accelerated diffusion samplers suffer as high dimensional dependencies as $\mathcal{O}\left(d^3\right)$ or $\mathcal{O}\left(d^2\right)$ (Li et al., 2024a;c) for target distributions with bounded support. This motivates us to explore the following intriguing question:

Question 2: While addressing Question 1 to relax the assumption from finite support to finite variance for possibly non-smooth distributions, can we achieve a lower dimensional dependency?

This paper will provide affirmative answers to both of the above questions.

1.1 Our Contributions

Our main contribution is to provide accelerated convergence results for a significantly wider range of distributions than those addressed in previous works (see Table 1 (particularly column 1) for a comparison). To this end, we design a new accelerated stochastic DDPM sampler and develop a novel analytical technique that characterizes its acceleration guarantees across this broader spectrum of distributions. Our detailed contributions are summarized as follows.

Broadening Target Distributions: Inspired by optimization methods, we design a new Hessian-based accelerated sampler for the stochastic diffusion processes. We show that our accelerated sampler achieves an accelerated convergence rate of $\mathcal{O}\left(d^{1.5}\min\{d,N\}^{1.5}/\varepsilon\right)$, $\mathcal{O}\left(d^{1.5}/\varepsilon\right)$, and $\mathcal{O}\left(d^{1.5}\log^{1.5}M/\varepsilon\right)$ respectively for Gaussian mixtures, any target distributions having finite variance (with early-stopping), and any target distributions having M-Lipschitz Hessian of log-densities. In particular, (i) for smoothness Q_0 that has p.d.f., the smoothness condition is only imposed on the log-density of Q_0, which is much less restrictive than that imposed on all Q_t's (Chen et al., 2023c; Huang et al., 2024a;b); (ii) for possibly non-smooth Q_0, we only require Q_0 to have finite variance for the early-stopped procedure, which is a much broader class of distributions than those having bounded support (Li et al., 2024a;c;d); (iii) we provide the first accelerated convergence result for Gaussian mixture Q_0's.[1]

For possibly non-smooth targets with bounded support, our sampler improves the dependency of the convergence rate on d by $\mathcal{O}\left(d^{1.5}\right)$ compared with previous accelerated diffusion samplers (Li et al., 2024a;c).

Novel Analysis Technique: We develop a novel technique for analyzing the accelerated DDPM process. Our approach features two new elements: (i) characterization of the error incurred at each discrete step of the reverse process using *tilting factor*; and (ii) analysis of the mean value of tilting factor via *Tweedie's formula* to handle power terms in the Taylor expansion. Such a technique enables us to (a) analyze more general distributions beyond those with restrictive distribution assumptions; (b) tightly identify the dominant term and reduce the dimensional dependency; and (c) handle the estimation error in accelerated samplers for both score and Hessian estimation. This analytical framework is different from the main previous theoretical techniques for analyzing the convergence of diffusion models: (a) the SDE-type analysis for regular diffusion samplers (Chen et al., 2023a; Benton et al., 2024a; Conforti et al., 2023), (b) any ODE-type analysis (Li et al., 2024d; Huang et al., 2024a; Gao & Zhu, 2024), and (c) the use of typical sets (Li et al., 2024a;c).

1.2 Related Works on Accelerated Sampling

Here, we focus on the related studies of accelerated samplers. Note that all of these works we discuss below, only except Chen et al. (2023c;e); Li et al. (2024c), are concurrent to or after ours based on their posting time on arXiv. In Appendix A, we provide a thorough summary of convergence analysis of standard samplers as well as other theoretical perspectives of diffusion models.

Accelerated Stochastic Samplers: In Li et al. (2024c), accelerated stochastic variants to the original DDPM sampler are proposed and analyzed, *when there is no estimation error*. In Li et al. (2024a), a new accelerated stochastic sampler are proposed by inserting intermediate sampling points along the diffusion path. Both algorithms are analyzed only when the target distribution has bounded support and suffer from large dimensional dependencies. In Huang et al. (2024b), the authors proposed the RTK-MALA and RTK-ULD algorithms which uses MCMC algorithms, such as the Metropolis-adjusted Langevin Algorithm or the Underdamped Langevin Dynamics, at each diffusion step. The analysis is performed under the assumption that all the scores of $\log q_t$'s are Lipschitz-smooth. In comparison, our work substantially broadens the set of target distributions to include those with unbounded support and with smooth log-density only imposed upon Q_0 with a completely different analytical technique. Our result also improves the dimensional dependencies of accelerated stochastic samplers in Li et al. (2024a;c) for distributions with bounded support.

Deterministic Samplers: Beyond stochastic samplers, another line of research to achieve an accelerated convergence rate is to sample from the corresponding probability flow ordinary differential equation (PF-ODE). Early work provided polynomial guarantees under rather restrictive Lipschitz conditions Chen et al. (2023e). Later in Chen et al. (2023c), an accelerated convergence rate was first derived with the DPUM sampler by mixing the deterministic predictor steps with stochastic corrector steps. The analysis was performed under the assumption of Lipschitz $\nabla \log q_t$'s and s_t's. Note that this assumption is relatively restrictive and hard to verify in practice. After that, for target distributions having bounded support, Li et al. (2024c) provided the first analysis of a purely deterministic sampler (along with an accelerated deterministic sampler), albeit with a high dimensional dependency. Recently, under strong assumptions on s_t's, Huang et al. (2024a) provided an accelerated rate using the p-th order Runge-Kutta time integrator for ODEs for those target distributions

[1]Although the technique in Huang et al. (2024a) may be applied to Gaussian mixtures, the authors do not provide explicit dependencies in their paper. Also, Huang et al. (2024a) is posted on arXiv after our first draft.

having bounded support. Specifically, for first-order Runge-Kutta methods, it is assumed that the first two orders of partial derivatives of s_t's are uniformly bounded in space and time, which implies Lipschitz-smoothness of s_t and its derivative along the entire sampling path. Most recently, Li et al. (2024d) obtained a linear convergence rate both in d and ε^{-1} using PF-ODEs as long as s_t's (and their derivatives) are well estimated. However, it is analyzed only on bounded-support targets. Beyond these works, further acceleration to deterministic samplers is sought in Li et al. (2024a;c) that gives the convergence rate of $\mathcal{O}(\varepsilon^{-1/2})$, which are still performed under bounded-support targets. In comparison, our work substantially broadens the target distributions to include those with unbounded support (yet with finite variance) while achieving an accelerated convergence rate.

2 Preliminaries of DDPM

In this section, we provide the background of the DDPM sampler (Ho et al., 2020).

2.1 Forward Process

Let $x_0 \in \mathbb{R}^d$ be the initial data, and let $x_t \in \mathbb{R}^d, t \in \{1, \ldots, T\}$ be the latent variables in the diffusion algorithm. Let Q_0 be the initial data distribution, and let Q_t be the marginal latent distribution at time t in the forward process, for all $1 \le t \le T$. In the forward process, white Gaussian noise is gradually added to the data: $x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}w_t$, $\forall t \in \{1, \ldots, T\}$, where $w_t \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$. Equivalently, this can be expressed as a conditional distribution at each time t:

$$Q_{t|t-1}(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I_d), \tag{1}$$

which means that under Q, $X_0 \to X_1 \to \cdots \to X_T$. Here $\beta_t \in (0,1)$ captures the "amount" of noise that is injected at time t, and β_t's are called the *noise schedule*. Define

$$\alpha_t := 1 - \beta_t, \quad \bar{\alpha}_t := \textstyle\prod_{i=1}^{t} \alpha_i, \quad 1 \le t \le T.$$

An immediate result by accumulating the steps is that

$$Q_{t|0}(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I_d), \tag{2}$$

or, written equivalently, $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\bar{w}_t$, $\forall t \in \{1, \ldots, T\}$, where $\bar{w}_t \sim \mathcal{N}(0, I_d)$ denotes the *aggregated* noise at time t. Intuitively, for large T, since $Q_{T|0} \approx \mathcal{N}(0, I_d)$ (which is independent of x_0), it is expected that $Q_T \approx \mathcal{N}(0, I_d)$ when T becomes large, as long as the variance under Q_0 is finite. Finally, since the conditional noises are Gaussian, each $Q_t (t \ge 1)$ is absolutely continuous w.r.t the Lebesgue measure. Let the corresponding p.d.f. of each Q_t be $q_t (t \ge 1)$. Similarly define $q_{t,t-1}$, $q_{t|t-1}$, and $q_{t-1|t}$ for $t \ge 1$. In case Q_0 is also absolutely continuous w.r.t. the Lebesgue measure, let q_0 be the corresponding p.d.f. of Q_0.

2.2 Regular Reverse Process

The goal of the reverse sampling process is to generate samples approximately from the data distribution Q_0. We first draw the latent variable at time T from a Gaussian distribution: $x_T \sim \mathcal{N}(0, I_d) =: P_T$. Then, to achieve effective sampling, each forward step is approximated by a reverse sampling step, in which the *mean* matches the posterior mean of $Q_{t-1|t}$. Define

$$\mu_t(x_t) := \tfrac{1}{\sqrt{\alpha_t}} \left(x_t + (1-\alpha_t)\nabla \log q_t(x_t)\right). \tag{3}$$

Here $\nabla \log q_t(x)$ is called the *score* of q_t, which can be estimated via a training process called score matching. At each time $t = T, T-1, \ldots, 1$, the *true* regular reverse process is defined as $x_{t-1} = \mu_t(x_t) + \sigma_t z$, where $z \sim \mathcal{N}(0, I_d)$. Two choices of σ_t^2 are commonly used in practice, where $\sigma_t^2 = 1 - \alpha_t$ or $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}(1-\alpha_t)$, and similar results are reported for these choices (Ho et al., 2020). Let P_t be the marginal distributions of x_t in the true regular reverse process, and let p_t be the corresponding p.d.f. of P_t w.r.t. the Lebesgue measure.

2.3 Metrics

In case where Q is absolutely continuous w.r.t. the Lebesgue measure, we are interested in measuring the mismatch between Q and P through the total-variation distance, defined as

$$\mathrm{TV}(Q, P) := \sup_{A \subseteq \mathcal{B}(\mathbb{R}^d)} |Q(A) - P(A)|$$

where $\mathcal{B}(\mathbb{R}^d)$ contains all Borel-measureable sets in $\mathbb{R}_d$. This metric is commonly used in prior theoretical studies (Chen et al., 2023a). From Pinsker's inequality, the total-variation (TV) distance is upper bounded as $\mathrm{TV}(Q,P)^2 \le \frac{1}{2}\mathrm{KL}(Q||P)$, where the KL divergence is defined as $\mathrm{KL}(Q||P) := \int \log \frac{\mathrm{d}Q}{\mathrm{d}P}\mathrm{d}Q \ge 0$. Thus, we control the KL divergence when Q is absolutely continuous w.r.t. P.

When q_0 does not exist (say, when Q_0 has point masses), we use the Wasserstein distance to measure the mismatch at $t=0$, namely $\mathrm{W}_2(Q_0,Q_1)$, which is a technique commonly adopted (Chen et al., 2023a; Benton et al., 2024a). The Wasserstein-2 distance is defined as $\mathrm{W}_2(Q_0,Q_1) := \sqrt{\min_{\Gamma\in\Pi(Q_0,Q_1)}\int_{\mathbb{R}^d\times\mathbb{R}^d}\|x-y\|^2\,\mathrm{d}\Gamma(x,y)}$, where $\Pi(Q_0,Q_1)$ is the set of all joint probability measures on $\mathbb{R}^d\times\mathbb{R}^d$ with marginal distributions Q_0 and Q_1, respectively.

3 Accelerated Diffusion Sampler

To generate samples from the data distribution Q_0, the idea of DDPM is to design a reverse process in which each reverse sampling step well approximates the corresponding forward step. Below, we propose a new **accelerated** sampler along with a new variance estimator, in which both the conditional *mean and variance* of the reverse process match the corresponding posterior quantities.

3.1 Accelerated Reverse Process

At each time $t=T,T-1,\dots,1$, define the true *accelerated* reverse process as $x_{t-1} = \mu_t(x_t) + \Sigma_t^{\frac{1}{2}}(x_t)z$, where μ_t is defined in (3), $z\sim\mathcal{N}(0,I_d)$, and (cf. Lemma 8)

$$\Sigma_t(x_t) := \tfrac{1-\alpha_t}{\alpha_t}\left(I_d + (1-\alpha_t)\nabla^2\log q_t(x_t)\right). \tag{4}$$

Let P'_t be the marginal distributions of x_t in the true accelerated reverse process, and let p'_t be the corresponding p.d.f.. Thus, the transition kernel can be written as $P'_{t-1|t} = \mathcal{N}(x_{t-1};\mu_t(x_t),\Sigma_t(x_t))$, and we let $P'_T := P_T = \mathcal{N}(0,I_d)$. When $(1-\alpha_t)$ is vanishing for large T, $\Sigma_t(x_t)\succ 0$ for all large T's, and thus the conditional Gaussian process is well-defined.[2] The above accelerated sampler has a close relationship to Ozaki's discretization method to approximate a continuous-time stochastic process (Ozaki, 1992; Shoji, 1998; Stramer & Tweedie, 1999).

In practice, one has no access to either $\nabla\log q_t$ or $\nabla^2\log q_t$. Thus, their estimates, denoted as s_t and H_t, are used. Define the *estimated* accelerated reverse process: $x_{t-1} = \widehat{\mu}_t(x_t) + \widehat{\Sigma}_t^{\frac{1}{2}}(x_t)z$, where

$$\widehat{\mu}_t(x_t) := x_t + (1-\alpha_t)s_t(x_t), \tag{5}$$

$$\widehat{\Sigma}_t(x_t) := \tfrac{1-\alpha_t}{\alpha_t}\left(I_d + (1-\alpha_t)H_t(x_t)\right). \tag{6}$$

Here, s_t can be obtained through score-matching (Song & Ermon, 2019). In Section 3.2, we propose an estimator for $\nabla^2\log q_t$, which we refer to as Hessian matching. Let $\widehat{P}'_t$ be the marginal distributions of x_t in the estimated reverse process with corresponding p.d.f. $\widehat{p}'_t$.

3.2 Hessian Matching Estimator for Acceleration

Below we provide a method to obtain $H_t(x)$, which estimates $\nabla^2\log q_t(x)$. Note that

$$\begin{aligned}\nabla^2\log q_t(x) &= \tfrac{\nabla^2 q_t(x)}{q_t(x)} - (\nabla\log q_t(x))(\nabla\log q_t(x))^\intercal \\ &= \left(\tfrac{\nabla^2 q_t(x)}{q_t(x)} + \tfrac{1}{1-\bar\alpha_t}I_d\right) - \tfrac{1}{1-\bar\alpha_t}I_d - (\nabla\log q_t(x))(\nabla\log q_t(x))^\intercal.\end{aligned} \tag{7}$$

Apart from the original score estimate, we require an additional Hessian estimate:

$$v_t(x) := \arg\min_{v_\theta:\mathbb{R}^d\to\mathbb{R}^{d\times d}} \mathbb{E}_{X_t\sim Q_t}\left\|v_\theta(X_t) - \left(\tfrac{\nabla^2 q_t(X_t)}{q_t(X_t)} + \tfrac{1}{1-\bar\alpha_t}I_d\right)\right\|_F^2.$$

In order to train for v_t, the following lemma provides an analogy to score matching, which we refer to as *Hessian matching*.

[2] More rigorously, we can project the matrices Σ_t and $\widehat{\Sigma}_t$ onto the space of positive-semi definite (PSD) matrices for those x_t's where either of these two matrices is not PSD. Since the measure of the events containing such bad x_t's decreases to zero, all theoretical results in this paper, which are derived in expectation, will not be affected.

Lemma 1. *With the forward process in* (1)*, we have*

$$\arg\min_{v_\theta:\mathbb{R}^d\to\mathbb{R}^{d\times d}} \mathbb{E}_{X_t\sim Q_t}\left\|v_\theta(X_t)-\left(\frac{\nabla^2 q_t(X_t)}{q_t(X_t)}+\frac{1}{1-\bar\alpha_t}I_d\right)\right\|_F^2$$
$$=\arg\min_{v_\theta:\mathbb{R}^d\to\mathbb{R}^{d\times d}} \mathbb{E}_{(X_0,\bar W_t)\sim Q_0\otimes\mathcal{N}(0,I_d)}\left\|v_\theta(\sqrt{\bar\alpha_t}X_0+\sqrt{1-\bar\alpha_t}\bar W_t)-\frac{1}{1-\bar\alpha_t}\bar W_t\bar W_t^\intercal\right\|_F^2.$$

With the Hessian estimate v_t using Lemma 1, from (7), an estimate for $\nabla^2\log q_t(x)$ is given by

$$H_t(x)=v_t(x)-\frac{1}{1-\bar\alpha_t}I_d-s_t(x)s_t^\intercal(x). \tag{8}$$

With the estimator of H_t in (8), the Hessian-based sampler using the $\widehat{\widehat{\Sigma}}_t$ later in (9) is the same as the accelerated stochastic sampler in Li et al. (2024c). Yet, our analysis is applicable when estimation errors exist, whereas in Li et al. (2024c) the estimators are assumed to be perfect for the accelerated sampler. In the literature, several other estimators have been proposed for higher order derivatives of $\log q_t(x)$ (Meng et al., 2021; Lu et al., 2022). In our paper, we proposed another method, the Hessian matching method, which can guarantee accurate Hessian estimations with extra computation resources. Our analysis is applicable to any estimator for H_t as long as Assumption 3 is satisfied.

4 Accelerated Convergence Bounds for Broader Targets

In this section, we provide convergence guarantees for the accelerated stochastic samplers for general Q_0. We will first establish our main result for smooth Q_0, and then extend it for more general (possibly non-smooth) Q_0. We will also provide a sketch of proof to describe key analysis techniques.

4.1 Technical Assumptions for Accelerated Sampler

We first provide the following four technical assumptions for the accelerated sampler.

Assumption 1 (Finite Second Moment)**.** There exists a constant $M_2<\infty$ (that does not depend on d and T) such that $\mathbb{E}_{X_0\sim Q_0}\|X_0\|^2\le M_2 d$.

Assumption 2 (Absolute Continuity)**.** Q_0 is absolutely continuous w.r.t. the Lebesgue measure, and thus q_0 exists. Also, suppose that q_0 is analytic [3] and that $q_0(x)>0$.

The above Assumptions 1 and 2 are commonly adopted in the literature (Chen et al., 2023a;d).

Assumption 3 (Score and Hessian Estimation Error)**.** The estimates s_t's and H_t's satisfy

$$\frac{1}{T}\sum_{t=1}^T\mathbb{E}_{X_t\sim Q_t}\|s_t(X_t)-\nabla\log q_t(X_t)\|^2\le\varepsilon^2=\tilde O(T^{-2}),$$
$$\frac{1}{T}\sum_{t=1}^T\mathbb{E}_{X_t\sim Q_t}\|H_t(X_t)-\nabla^2\log q_t(X_t)\|_F^2\le\varepsilon_H^2=\tilde O(T^{-1}).$$

Also, suppose that H_t satisfies $\sup_{\ell\ge1}\left(\mathbb{E}_{X_t\sim Q_t}\|H_t(X_t)\|^\ell\right)^{1/\ell}=\tilde O(1)$.

The above assumption (Assumption 3) describes the estimation error for both the score and Hessian. In particular, compared with regular samplers, the score function needs to be estimated at a higher accuracy in order to achieve acceleration. Such higher accuracy is also required in previous analyses of ODE samplers (e.g., Li et al. (2024a;d)). The regularity condition on H_t can be satisfied, for example, when $\|H_t\|$ is bounded as $\tilde O(1)$. As another example, it suffices that $\|H_t(x)\|$ has a polynomial upper bound in x when Q_t is sub-exponential. In Lemma 2 (in Appendix C), we provide sufficient conditions such that the H_t in (8) satisfies Assumption 3.

Assumption 4 (Regular Partial Derivatives)**.** For all $t\ge1$, $\ell\ge1$, and $\boldsymbol{a}\in[d]^p$ such that $|\boldsymbol{a}|=p\ge1$,

$$\mathbb{E}_{X_t\sim Q_t}\left|\partial_{\boldsymbol{a}}^p\log q_t(X_t)\right|^\ell=O(1),\quad \mathbb{E}_{X_t\sim Q_t}\left|\partial_{\boldsymbol{a}}^p\log q_{t-1}(\mu_t(X_t))\right|^\ell=O(1).$$

When q_0 does not exist, this is required only for $t\ge2$.

The above regularity assumption (Assumption 4) on the partial derivatives is needed for our analysis based on Taylor expansion.[4] It is rather soft, and it can be verified on the following two common cases: (1) when Q_0 has finite variance, and (2) when Q_0 is Gaussian mixture (see Section 5). Case 1 clearly covers a broad set of target distributions of practical interest, such as images, and many theoretical studies of diffusion models have been specially focused on such a distribution (Li et al., 2024a;c). Case 2 has also been well studied for diffusion models (Chen et al., 2024; Gatmiry et al., 2024).

[3] Here a function is analytic if its Taylor series converges to the functional value at each point in the domain.

[4] In the Appendix, we have provided the more general Assumption 5 under which Theorem 1 would hold.

4.2 Accelerated Convergence Bounds

We first define a new noise schedule as follows, which will be useful for acceleration.

Definition 1 (Noise Schedule for Acceleration)**.** For large T's, the step-size α_t satisfies that

$$1-\alpha_t \lesssim \frac{\log T}{T},\ \forall t \in \{1,\dots,T\}, \quad \bar{\alpha}_T = \prod_{t=1}^T \alpha_t = o\left(T^{-2}\right).$$

When q_0 does not exist, the upper bound on $1-\alpha_t$ is only required for $t \geq 2$.

In Definition 1, the upper bound on $1-\alpha_t$ requires that α_t is large enough to control the reverse-step error, while the upper bound on $\bar{\alpha}_T$ requires that α_t is small enough to control the initialization error. An example of α_t that satisfies Definition 1 is the constant step-size: $1-\alpha_t \equiv \frac{c\log T}{T},\ \forall t \geq 1$ with $c > 2$. Then, $\bar{\alpha}_T = \left(1-\frac{c\log T}{T}\right)^T = \exp\left(T\log\left(1-\frac{c\log T}{T}\right)\right) = O\left(e^{T\frac{-c\log T}{T}}\right) = o\left(T^{-2}\right)$. Thus, such α_t satisfies Definition 1.

The following theorem provides the *first* convergence result for accelerated diffusion samplers for general smooth target distributions that have *finite second moment* (along with some mild regularity conditions). The complete proof is given in Appendix D.

Theorem 1 (Accelerated Sampler for Smooth Q_0)**.** *Under Assumptions 1 to 4, with the α_t satisfying Definition 1, we have*

$$\begin{aligned}\mathrm{KL}(Q_0||\widehat{P}_0') \lesssim & (\log T)\varepsilon^2 + \frac{\log^2 T}{T}\varepsilon_H^2 \\ & + \sum_{t=1}^T (1-\alpha_t)^3 \mathbb{E}_{X_t\sim Q_t} \sum_{i,j,k=1}^d \partial_{ijk}^3 \log q_{t-1}(\mu_t(X_t))\partial_{ijk}^3 \log q_t(X_t).\end{aligned}$$

Theorem 1 characterizes the convergence in terms of KL divergence (and thus TV distance) for smooth (possibly unbounded) Q_0. The bound in Theorem 1 will be further instantiated with explicit dependency on system parameters for example distributions Q_0 in Section 5. To further explain the upper bound in Theorem 1, the first two terms arise from the score and Hessian estimation error, and the last term captures the errors accumulated during the reverse steps over $t = T,\dots,1$, which can be further bounded by $\tilde{O}(T^{-2})$ under Assumption 4 (cf. (52)). Thus, when ε_H^2 satisfies Assumption 3, the upper bound in Theorem 1 can be more explicitly characterized w.r.t. T as $\mathrm{KL}(Q_0||\widehat{P}_0) \lesssim \tilde{O}(T^{-2}) + (\log T)\varepsilon^2$ (where the dependency on d will be explicitly characterized for specific distributions in Section 5). Thus, in order to achieve $\mathcal{O}(\varepsilon^2)$ error in KL divergence, the number of steps required is $\mathcal{O}(\varepsilon^{-1})$. This improves the dependency of the convergence rate on ε of the regular sampler by a factor of $\mathcal{O}(\varepsilon^{-1})$.

We next extend Theorem 1 for smooth Q_0 to general Q_0 that can be possibly non-smooth and hence the density function q_0 does not exist. Such distributions occur often in practice; for example, when Q_0 has a discrete support such as for images, or when Q_0 is supported on a low-dimensional manifold. For non-smooth Q_0, its one-step perturbation Q_1 does have a p.d.f. q_1, which is further analytic (Lemma 6). This enables us to apply Theorem 1 on Q_1 to obtain the following convergence bound. Also, we use the Wasserstein distance to measure the perturbation between Q_0 and Q_1 (Chen et al., 2023d;a; Lee et al., 2023).

Corollary 1 (General (possibly non-smooth) Q_0)**.** *Under Assumptions 1, 3 and 4, if the noise schedule satisfies Definition 1 at $t \geq 2$, the distribution $\widehat{P}_1'$ satisfies*

$$\begin{aligned}\mathrm{KL}(Q_1||\widehat{P}_1') \lesssim & (\log T)\varepsilon^2 + \frac{\log^2 T}{T}\varepsilon_H^2 \\ & + \sum_{t=2}^T (1-\alpha_t)^3 \mathbb{E}_{X_t\sim Q_t} \sum_{i,j,k=1}^d \partial_{ijk}^3 \log q_{t-1}(\mu_t(X_t))\partial_{ijk}^3 \log q_t(X_t),\end{aligned}$$

where Q_1 is such that $\mathrm{W}_2(Q_0,Q_1)^2 \lesssim (1-\alpha_1)d$.

In particular, Corollary 1 applies to any general target distribution when the second moment is finite.

4.3 Proof Sketch of Theorem 1

We next provide a proof sketch of Theorem 1 to describe the idea of our analysis approach. The full proof is provided in Appendix D. Our approach is very different from previous SDE-type approaches, which invoke Fokker-Planck equation to express the evolution of p.d.f. and use Girsanov's Theorem to bound the divergence, both along the *continuous* diffusion path. In comparison, we develop a novel Bayesian approach based on tilting factor representation and Tweedie's formula to handle power terms, which is applicable to a much wider class of target distributions, including those having

infinite support. In particular, compared with Li et al. (2024a;c;d), our approach does not assume that the target distribution has finite support.

To begin, we decompose the total error as

$$\mathrm{KL}(Q_0||\widehat{P}'_0) \le \underbrace{\mathbb{E}_{X_T\sim Q_T}\left[\log\frac{q_T(X_T)}{p'_T(X_T)}\right]}_{\text{initialization error}} + \underbrace{\sum_{t=1}^T \mathbb{E}_{X_t,X_{t-1}\sim Q_{t,t-1}}\left[\log\frac{p'_{t-1|t}(X_{t-1}|X_t)}{\widehat{p}'_{t-1|t}(X_{t-1}|X_t)}\right]}_{\text{estimation error}} + \underbrace{\sum_{t=1}^T \mathbb{E}_{X_t,X_{t-1}\sim Q_{t,t-1}}\left[\log\frac{q_{t-1|t}(X_{t-1}|X_t)}{p'_{t-1|t}(X_{t-1}|X_t)}\right]}_{\text{reverse-step error}}.$$

The initialization error can be bounded easily (Lemma 3). Below we focus on the remaining two terms in five steps.

Step 1: Bounding estimation error (Lemma 4). At each time $t = 1, \dots, T$, rather than upper-bounding via typical sets as in Li et al. (2024c), we directly evaluate the expected value of $\log(p'_{t-1|t}(x_{t-1}|x_t)/\widehat{p}'_{t-1|t}(x_{t-1}|x_t))$. This is straightforward since $P'_{t-1|t}$ and $\widehat{P}'_{t-1|t}$ are Gaussian. We then use Taylor expansion for the $\log\det(\cdot)$ function and the matrix inverse to identify the dominant-order terms under the mismatched variance.

Step 2: Tilting factor expression of log-likelihood ratio (Lemmas 5 and 6 and Equation (20)). With Bayes' rule, we show that $q_{t-1|t}$ is an exponentially tilted form of $p'_{t-1|t}$ with tilting factor:

$$\zeta'_{t,t-1} = (\nabla \log q_{t-1}(\mu_t) - \sqrt{\alpha_t}\nabla \log q_t(x_t))^\intercal (x_{t-1}-\mu_t) + \tfrac{1}{2}(x_{t-1}-\mu_t)^\intercal\left(\nabla^2 \log q_{t-1}(\mu_t) - \tfrac{\alpha_t}{1-\alpha_t}B_t(x_t)\right)(x_{t-1}-\mu_t) + \textstyle\sum_{p=3}^\infty T_p(\log q_{t-1}, x_{t-1}, \mu_t).$$

where $B_t(x_t)$ describes the correction due to the modified variance for acceleration (see (14)), and $T_p(f, x, \mu)$ is the p-th order Taylor power term of function f around $x = \mu$. With this tilting factor, we can upper-bound the reverse-step error as, for each fixed x_t,

$$\mathbb{E}_{X_{t-1},X_t\sim Q_{t-1,t}}\left[\log\frac{q_{t-1|t}(X_{t-1}|x_t)}{p'_{t-1|t}(X_{t-1}|x_t)}\right] \le \mathbb{E}_{X_t,X_{t-1}\sim Q_{t,t-1}}[\zeta'_{t,t-1}] - \mathbb{E}_{X_t\sim Q_t,X_{t-1}\sim P'_{t-1|t}}[\zeta'_{t,t-1}].$$

For regular DDPMs, there is no control for the variance of the reverse sampling process, and thus $B_t(x_t) \equiv 0$. In this case, the dominating rate is determined by the expected values of T_2. With the variance correction in our accelerated sampler, the corresponding $B_t(x_t)$ enables us to cancel out the second-order Taylor term (see Lemma 11). As a result, the rate-determining term becomes the expected values of T_3, which decays faster. Thus, the acceleration is achieved.

Step 3: Explicit expression for $\mathbb{E}_{X_t\sim Q_t,X_{t-1}\sim P'_{t-1|t}}[\zeta'_{t,t-1}]$ (Lemma 7). Given the Taylor expansion of $\zeta'_{t,t-1}$, this step can be reduced to calculating the expected values of the power terms, which are the Gaussian centralized moments. They are calculated using the classical Isserlis's Theorem.

Step 4: Explicit expression for $\mathbb{E}_{X_t,X_{t-1}\sim Q_{t,t-1}}[\zeta'_{t,t-1}]$ (Lemmas 8 to 10). While $Q_{t|t-1}$ is Gaussian, $Q_{t-1|t}$ is not Gaussian in general, rendering the calculation of all moments non-trivial. To calculate posterior moments, we extend Tweedie's formula (Efron, 2011) in a non-trivial way. Whereas the original Tweedie's formula provides an explicit expression for the posterior mean for Gaussian perturbed observations, we explicitly calculate the first six centralized posterior moments and provide the asymptotic order of all higher-order moments, drawing techniques from combinatorics. The results also justify the expressions of μ_t and Σ_t in (3) and (4).

Step 5: Bounding reverse-step error (Lemma 11) In order to employ the moment results for Taylor expansion, we guarantee that it is valid to change the limit (in the Taylor expansion) and the expectation operator. Finally, substituting the calculated moments into $\mathbb{E}_{X_t,X_{t-1}\sim Q_{t,t-1}}[\zeta'_{t,t-1}] - \mathbb{E}_{X_t\sim Q_t,X_{t-1}\sim P'_{t-1|t}}[\zeta'_{t,t-1}]$ and noting that higher-order partial derivatives do not affect the rate (by Assumption 4), we can determine the dominating term and obtain the desirable result.

5 Example Q_0's: Accelerated Convergence Rate with Explicit Parameter Dependency

Now, we specialize Theorem 1 and Corollary 1 to several interesting distribution classes, for which convergence bounds with explicit dependency on system parameters can be derived. The key is to locate the dependency in the dominating terms in the reverse-step error.

5.1 GAUSSIAN MIXTURE Q_0

We first investigate the case where Q_0 is Gaussian mixture. This is a rich class of distributions with strong approximation power (Bacharoglou, 2010; Diakonikolas et al., 2017). The following theorem establishes the first accelerated convergence result with explicit dimensional dependencies for such a distribution class.

Theorem 2 (Accelerated Sampler for Gaussian Mixture Q_0)**.** *Suppose that Q_0 is Gaussian mixture, whose p.d.f. is given by $q_0(x_0) = \sum_{n=1}^N \pi_n q_{0,n}(x_0)$, where $q_{0,n}$ is the p.d.f. of $\mathcal{N}(\mu_{0,n}, \Sigma_{0,n})$ and $\pi_n \in [0,1]$ is the mixing coefficient where $\sum_{n=1}^N \pi_n = 1$. Under Assumption 3, if the α_t satisfies Definition 1, we have*

$$\mathrm{KL}(Q_0||\widehat{P}_0') \lesssim \frac{d^3 \min\{d,N\}^3 \log^3 T}{T^2} + (\log T)\varepsilon^2 + \frac{\log^2 T}{T}\varepsilon_H^2.$$

Therefore, for any Gaussian mixture target Q_0 with $N \le d$, it takes the accelerated algorithm $\mathcal{O}\left(d^{1.5}N^{1.5}/\varepsilon\right)$ steps to reach convergence under accurate score and Hessian estimation. This is the first result for accelerated DDPM samplers to achieve an accelerated convergence rate for Gaussian mixture targets under score and Hessian estimation error. Compared with the results for regular samplers, the number of convergence steps improves by a factor of $\mathcal{O}(\varepsilon^{-1})$.

The proof of Theorem 2 is non-trivial because in order to show that Assumption 4 holds for Gaussian mixture distributions with any α_t according to Definition 1, it is generally difficult to evaluate and provide an upper bound for *all orders* of partial derivatives of the logarithm of a mixture density. To this end, we employ the multivariate Faá di Bruno's formula (Constantine & Savits, 1996) to develop an explicit bound (Lemmas 13 and 14).

Below we numerically evaluate the performance of our Hessian-accelerated DDPM when Q_0 is Gaussian mixture. The original accelerator requires calculating the square-root matrix of $\widehat{\Sigma}_t$ (see (4)), which might be computational burdensome. Below, we propose an approximated Hessian-based accelerated sampler, where $\widehat{\mu}_t$ is still defined in (5) and $\widehat{\Sigma}_t$ is replaced by $\widehat{\tilde{\Sigma}}_t(x_t)$ where

$$\tilde{\Sigma}_t(x_t) := \frac{1-\alpha_t}{\alpha_t}\left(I_d + \frac{1-\alpha_t}{2}\nabla \log q_t(x_t)\right)^2, \ \widehat{\tilde{\Sigma}}_t(x_t) := \frac{1-\alpha_t}{\alpha_t}\left(I_d + \frac{1-\alpha_t}{2}H_t(x_t)\right)^2. \tag{9}$$

With a similar tilting-factor analysis as in Theorem 1, we can verify that the approximated sampler still achieves an accelerated convergence rate (see Corollaries 2 and 3 and Remark 3).

In Figure 1, we compare the following four accelerated samplers: (1) the regular DDPM sampler (in blue); (2) our Hessian-accelerated sampler (in red); (3) the accelerated stochastic sampler in Li et al. (2024a) (in cyan); and (4) the deterministic sampler using PF-ODE, which is analyzed in Li et al. (2024c;d); Huang et al. (2024a). Here $N = 4$ and $d = 4$. The performance is averaged over 30 different trials. In a single trial, 200000 samples are used to estimate the KL divergence. The α_t in (10) is used with $c = 4$ and $\delta = 0.001$. From the comparison, it is observed that our Hessian-based sampler achieves the best convergence (at similar computation levels) in non-asymptotic regimes.

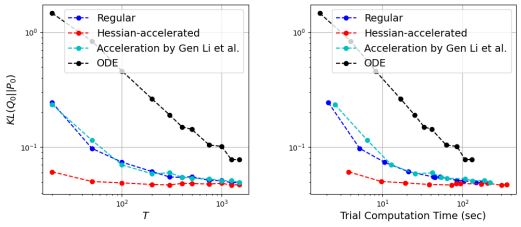

Figure 1: Comparison of different accelerated samplers for Gaussian mixture Q_0's. The x-axes are the number of steps (left) and the computation time of a trial (right), respectively.

5.2 FINITE VARIANCE Q_0 WITH EARLY-STOPPING

Next, we specialize Corollary 1 to a special noise schedule, first proposed in Li et al. (2024c):

$$1 - \alpha_t = \frac{c\log T}{T}\min\left\{\delta\left(1 + \frac{c\log T}{T}\right)^t, 1\right\}, \ \forall 2 \le t \le T, \tag{10}$$

and $1-\alpha_1=\delta$. Here c and δ satisfy that $c>2$ and $\delta e^c>1$. Intuitively, δ characterizes the amount of perturbation between Q_1 and Q_0 (Lemma 12). Note that any noise schedule satisfying the above condition also satisfies Definition 1 at $t\geq 2$ (see (49)), and hence Corollary 1 still holds here.

Theorem 3 (Accelerated Sampler for Q_0 with Finite Variance)**.** *Under Assumptions 1 and 3, using the α_t defined in* (10) *with $c>2$ and $c\asymp\log(1/\delta)$, we have*

$$\mathrm{KL}(Q_1||\widehat{P}_1')\lesssim\frac{d^3\log^3(1/\delta)\log^3 T}{T^2}+(\log T)\varepsilon^2+\frac{\log^2 T}{T}\varepsilon_H^2,$$

where Q_1 is such that $\mathrm{W}_2(Q_0,Q_1)^2\lesssim\delta d$.

Theorem 3 indicates that for any Q_0 having *finite variance*, it takes the accelerated algorithm $\mathcal{O}\left(d^{1.5}\log^{1.5}(1/\delta)/\varepsilon\right)$ steps to approximate an early-stopped data distribution Q_1 within $\mathcal{O}(\varepsilon^2)$ error in KL divergence (or $\mathcal{O}(\varepsilon)$ in TV distance). For early-stopped procedures, this theorem significantly relaxes the previous assumption on the target distribution that requires Q_0 to have bounded support (Li et al., 2024a;c; Huang et al., 2024a; Li et al., 2024d). Compared to previous accelerated diffusion samplers for bounded-support targets (Li et al., 2024a;c), our number of convergence steps to achieve ε-TV distance has improved by a factor of $\mathcal{O}(d^{1.5})$.

The proof of Theorem 3 involves the following novel elements. (i) Verifying Assumption 4 requires evaluating and providing an upper bound for *all orders* of partial derivatives of the logarithm of a *continuous* mixture density. Differently from the case of Gaussian (discrete) mixture, here we can only have an upper bound in expectation (i.e., in $\mathcal{L}^p(Q_t)$) (Lemma 15). (ii) The second half of Assumption 4 requires an upper bound for the one-step perturbed score, which can be shown using the change-of-variable formula and the data processing inequality for large T (Lemmas 16 and 17).

5.3 Q_0 WITH LIPSCHITZ HESSIAN LOG-DENSITY

With the α_t in (10), we derive a convergence result when only the log-density of Q_0 is smooth.

Theorem 4 (Accelerated Sampler for Smooth Hessian Log-Density)**.** *Suppose that $\nabla^2\log q_0(x)$ is 2-norm M-Lipschitz. This means that $\exists M>0$ such that*

$$\left\|\nabla^2\log q_0(x)-\nabla^2\log q_0(y)\right\|\leq M\left\|x-y\right\|,\quad\forall x,y\in\mathbb{R}^d.$$

Then, under Assumptions 1 and 3, using the α_t in (10) *with $\delta=1/(M^{\frac{2}{3}}T^{\frac{3}{2}})$ and $c\geq\log(M^{\frac{2}{3}}T^{\frac{3}{2}})$, we have*

$$\mathrm{KL}(Q_0||\widehat{P}_0')\lesssim\frac{d^3(\log^3 M+\log^3 T)\log^3 T}{T^2}+(\log T)\varepsilon^2+\frac{\log^2 T}{T}\varepsilon_H^2.$$

We also provide an accelerated convergence result with linear d dependency when all the $\nabla^2\log q_t(x)$ $(t\geq 0)$ are 2-norm M-Lipschitz (see Theorem 5 in Appendix G.3).

Theorem 4 provides us with the *first* accelerated DDPM result with only a smoothness constraint on $\log q_0$, under the score and Hessian estimation error. In words, in order to reach $\mathcal{O}(\varepsilon)$ TV-distance when $\varepsilon_H^2/T\lesssim\varepsilon^2$, the number of steps needed under Lipschitz-Hessian Q_0's is $\mathcal{O}(d^{1.5}\log^{1.5}M/\varepsilon)$. This is different from Chen et al. (2023c); Huang et al. (2024a;b) in which some smoothness condition is imposed on all $\nabla\log q_t$'s (or s_t's or both). Compared with Theorem 3, this upper bound in Theorem 4 is directly over $\mathrm{KL}(Q_0||\widehat{P}_0')$ instead of for some early-stopped distribution. Our results provide new contributions that complement existing studies by exploring different assumptions of distributions, which enriches the existing set of distributions studied in the literature.

Our analysis is significantly different from that in (Chen et al., 2023a, Theorem 5). There, the Poincaré inequality is key to guarantee that the Lipschitz smoothness in $\nabla\log q_0$ is preserved when δ is small, but this inequality may not hold in our case with smoothness only in $\nabla^2\log q_0$. Instead, with smooth $\nabla^2\log q_0$, we expand the tilting factor only to its third-order Taylor polynomial and directly provide an upper bound with techniques used in proving Theorems 3 and 5.

6 CONCLUSION

In this paper, we have provided accelerated convergence guarantees for a much larger set of target distributions than in prior literature, including both smooth Q_0 and general Q_0 with early-stopping. The accelerated rates are achieved with a new accelerated Hessian-based DDPM sampler using a novel analysis technique. One future direction is to further shrink the d dependency for general Q_0. It is also interesting to investigate other acceleration schemes to further improve diffusion samplers.

REFERENCES

Athanassia G. Bacharoglou. Approximation of probability distributions by convex mixtures of gaussian measures. *Proceedings of the American Mathematical Society*, 138(7):2619–2628, 2010.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, 2024a.

Joe Benton, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods. *Transactions on Machine Learning Research*, 2024b.

Stefano Bruno, Ying Zhang, Dong-Young Lim, Ömer Deniz Akyildiz, and Sotirios Sabanis. On diffusion-based generative models and their error bounds: The log-concave case with full convergence estimates. *arXiv preprint arXiv:2311.13584*, 2023.

Yu Cao, Jingrun Chen, Yixin Luo, and Xiang ZHOU. Exploring the optimal choice for generative processes in diffusion models: Ordinary vs stochastic differential equations. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Jinyuan Chang, Zhao Ding, Yuling Jiao, Ruoxuan Li, and Jerry Zhijian Yang. Deep conditional generative learning: Model and error analysis. *arXiv preprint arXiv:2402.01460*, 2024.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: user-friendly bounds under minimal smoothness assumptions. In *Proceedings of the 40th International Conference on Machine Learning*, 2023a.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *Proceedings of the 40th International Conference on Machine Learning*, 2023b.

Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ODE is provably fast. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023c.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023d.

Sitan Chen, Giannis Daras, and Alexandros G. Dimakis. Restoration-degradation beyond linear diffusions: a non-asymptotic analysis for ddim-type samplers. In *Proceedings of the 40th International Conference on Machine Learning*, 2023e.

Sitan Chen, Vasilis Kontonis, and Kulin Shah. Learning general gaussian mixtures with efficient score matching. *arXiv preprint arXiv:2404.18893*, 2024.

Xiuyuan Cheng, Jianfeng Lu, Yixin Tan, and Yao Xie. Convergence of flow-based generative models via proximal gradient descent in wasserstein space. *arXiv preprint arXiv:2310.17582*, 2023.

Frank Cole and Yulong Lu. Score-based generative models break the curse of dimensionality in learning a family of sub-gaussian distributions. In *The Twelfth International Conference on Learning Representations*, 2024.

Giovanni Conforti, Alain Durmus, and Marta Gentiloni Silveri. Score diffusion models without early stopping: finite fisher information is all you need. *arXiv preprint arXiv:2308.12240*, 2023.

G Constantine and T Savits. A multivariate faa di bruno formula with applications. *Transactions of the American Mathematical Society*, 348(2):503–520, 1996.

F. Croitoru, V. Hondru, R. Ionescu, and M. Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 45(9):10850–10869, Sep 2023.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17695–17709, 2021.

Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 73–84, 2017.

Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Genie: higher-order denoising diffusion solvers. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Xuefeng Gao and Lingjiong Zhu. Convergence analysis for general probability flow odes of diffusion models in wasserstein distances. *arXiv preprint arXiv:2401.17958*, 2024.

Xuefeng Gao, Hoang M. Nguyen, and Lingjiong Zhu. Wasserstein convergence guarantees for a general class of score-based generative models. *arXiv preprint arXiv:2311.11003*, 2023.

Yuan Gao, Jian Huang, Yuling Jiao, and Shurong Zheng. Convergence of continuous normalizing flows for learning probability distributions. *arXiv preprint arXiv:2404.00551*, 2024.

Khashayar Gatmiry, Jonathan Kelner, and Holden Lee. Learning mixtures of gaussians using diffusion models. *arXiv preprint arXiv:2404.18869*, 2024.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Daniel Zhengyu Huang, Jiaoyang Huang, and Zhengjiang Lin. Convergence analysis of probability flow ode for score-based generative models. *arXiv preprint arXiv:2404.09730*, 2024a.

Xunpeng Huang, Difan Zou, Hanze Dong, Yi Zhang, Yi-An Ma, and Tong Zhang. Reverse transition kernel: A flexible framework to accelerate diffusion inference. *2405.16387*, 2024b.

Yuling Jiao, Yanming Lai, Yang Wang, and Bokai Yan. Convergence analysis of flow matching in latent space with transformers. *arXiv preprint arXiv:2404.02538*, 2024.

Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88, August 2023.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. In *Advances in Neural Information Processing Systems*, 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201, pp. 946–985, 2023.

Gen Li, Yu Huang, Timofey Efimov, Yuting Wei, Yuejie Chi, and Yuxin Chen. Accelerating convergence of score-based diffusion models, provably. *arXiv preprint arXiv:2403.03852*, 2024a.

Gen Li, Zhihan Huang, and Yuting Wei. Towards a mathematical theory for consistency training in diffusion models. *arXiv preprint arXiv:2402.07802*, 2024b.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards faster non-asymptotic convergence for diffusion-based generative models. In *The Twelfth International Conference on Learning Representations*, 2024c.

Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. A sharp convergence theory for the probability flow odes of diffusion models. *arXiv preprint arXiv:2408.02320*, 2024d.

Puheng Li, Zhong Li, Huishuai Zhang, and Jiang Bian. On the generalization properties of diffusion models. *arXiv preprint arXiv:2311.01797*, 2024e.

Cheng Lu, Kaiwen Zheng, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Maximum likelihood training for score-based diffusion odes by high-order denoising score matching. In *International Conference on Machine Learning*, 2022.

Junlong Lyu, Zhitang Chen, and Shoubo Feng. Sampling is as easy as keeping the consistency: convergence guarantee for consistency models, 2024.

Song Mei and Yuchen Wu. Deep networks as denoising algorithms: Sample-efficient learning of diffusion models in high-dimensional graphical models. *arXiv preprint arXiv:2309.11420*, 2023.

Chenlin Meng, Yang Song, Wenzhe Li, and Stefano Ermon. Estimating high order gradients of the data distribution by denoising. In *Advances in Neural Information Processing Systems*, 2021.

Pierre Moulin and Venugopal V. Veeravalli. *Statistical Inference for Engineers and Data Scientists*. Cambridge University Press, Cambridge, UK, 2018.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.

Tohru Ozaki. A bridge between nonlinear times series models and nonlinear stochastic dynamical systems: A local linearization approach. *Statistica Sinica*, 2(1):113–135, 1992.

Francesco Pedrotti, Jan Maas, and Marco Mondelli. Improved convergence of score-based diffusion models via prediction-correction. *arXiv preprint arXiv:2305.14164*, 2023.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 1530–1538, 2015.

Kulin Shah, Sitan Chen, and Adam Klivans. Learning mixtures of gaussians using the DDPM objective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Isao Shoji. Approximation of continuous time stochastic processes by a local linearization method. *Mathematics of Computation*, 67(221):287–298, 1998.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 2256–2265, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

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

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

O. Stramer and R. L. Tweedie. Langevin-type models ii: Self-targeting candidates for mcmc algorithms. *Methodology And Computing In Applied Probability*, 1(3):307–328, 1999.

Andre Wibisono, Yihong Wu, and Kaylee Yingxi Yang. Optimal score estimation via empirical bayes smoothing. *arXiv preprint arXiv:2402.07747*, 2024.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4), Nov 2023.

Kaihong Zhang, Heqi Yin, Feng Liang, and Jingbo Liu. Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. *arXiv preprint arXiv:2402.15602*, 2024.