THEORY ON SCORE-MISMATCHED DIFFUSION MODELS AND ZERO-SHOT CONDITIONAL SAMPLERS

Anonymous authors
Paper under double-blind review

ABSTRACT

The denoising diffusion model has recently emerged as a powerful generative technique, capable of transforming noise into meaningful data. While theoretical convergence guarantees for diffusion models are well established when the target distribution aligns with the training distribution, practical scenarios often present mismatches. One common case is in zero-shot conditional diffusion sampling, where the target conditional distribution is different from the (unconditional) training distribution. These score-mismatched diffusion models remain largely unexplored from a theoretical perspective. In this paper, we present the first performance guarantee with explicit dimensional dependencies for general score-mismatched diffusion samplers, focusing on target distributions with finite second moments. We show that score mismatches result in an asymptotic distributional bias between the target and sampling distributions, proportional to the accumulated mismatch between the target and training distributions. This result can be directly applied to zero-shot conditional samplers for any conditional model, irrespective of measurement noise. Interestingly, the derived convergence upper bound offers useful guidance for designing a novel bias-optimal zero-shot sampler in linear conditional models that minimizes the asymptotic bias. For such bias-optimal samplers, we further establish convergence guarantees with explicit dependencies on dimension and conditioning, applied to several interesting target distributions, including those with bounded support and Gaussian mixtures. Our findings are supported by numerical studies.

1 INTRODUCTION

Generative modeling stands as a cornerstone in deep learning, with the goal of producing samples whose distribution emulates that of the training data. Traditional approaches encompass variational autoencoders (VAE) (Kingma & Welling, 2022), generative adversarial networks (GANs) (Goodfellow et al., 2014), normalizing flows (Rezende & Mohamed, 2015), and others. Recently, diffusion models, especially the denoising diffusion probabilistic models (DDPMs) (Sohl-Dickstein et al., 2015; Ho et al., 2020), have emerged as particularly compelling generative models, gaining widespread acclaim for their stable and cutting-edge performance across various tasks, such as image and video generation (Ramesh et al., 2022; Rombach et al., 2022).

In ideal situations, the training and target distributions of generative models match each other. However, this often does not hold in practice, where distributional mismatch between the training and target distributions can occur due to various reasons such as possible privacy constraints, need for computational efficiency, and knowledge gap between training and sampling processes. Specifically for diffusion models, such mismatches exhibit between the scores obtained from the training data and the scores of the target distribution from which we want to generate samples. One common scenario that existing studies primarily focus on is **conditional** diffusion models in image generation tasks (see Croitoru et al. (2023); Li et al. (2023); Moser et al. (2024) for surveys of diffusion models in computer vision). Different from unconditional image generation, conditional image samplers aim to generate images that are consistent with the given information, either be a text-prompt (as in text-to-image synthesis) or a sub-image (as in image super-resolution). For example, in image super-resolution, given the input of a low-resolution image, the goal is not to generate some arbitrary high-resolution image but the one whose corresponding low-resolution part matches the given input. Here the diffusion models are well-trained on the *unconditional* distribution of high-resolution images, whereas the target distribution is the *conditional* distribution given the low-resolution input. If one uses these well-trained unconditional scores to generate conditional samples, there will be a mismatch at each step of the sampling process.

One class of methods to tackle the conditional sampling problem is to include *extra-guided training*, where a modified score function is trained with the *extra* knowledge of the conditioning information (Dhariwal & Nichol, 2021; Ho & Salimans, 2022). On the theory side, several recent works (Yuan et al., 2023; Wu et al., 2024; Fu et al., 2024) provided the performance guarantee for such conditional diffusion samplers, where a score guidance is obtained through *extra* training based on the conditional information. However, the additional guided training in these samplers requires *extra* computations and needs to be conducted for every image conditioning, which may not be efficient in practice.

Alternatively, *zero-shot* conditional image samplers arise as a prevalent approach (e.g., Choi et al. (2021); Chung et al. (2022b;a; 2023); Wang et al. (2023); Song et al. (2023a); Fei et al. (2023)) for *training-free* conditional generation given well-trained unconditional scores. For each conditioned image, zero-shot samplers require no additional training to modify the scores. Instead, they adjust the scores during sampling by calculating rectified scores based on conditional information to mitigate the mismatch between the oracle conditional scores and the approximated ones.[1] Despite their empirical promise, theoretical guarantee on these zero-shot samplers is largely unexplored. In Gupta et al. (2024), the authors provided a super-polynomial lower bound for zero-shot sampling as a converse result. In Xu & Chi (2024), the authors proposed and analyzed a *plug-and-play* conditional sampler. However, their analysis relies on the properties of the Markov transition kernel specific to their plug-and-play model, which does not appear to be applicable to several widely used zero-shot samplers, such as Come-Closer-Diffuse-Faster (CCDF) (Chung et al., 2022b) and Denoising Diffusion Null-space Model (DDNM) (Wang et al., 2023). Therefore, there is a need to provide the performance guarantee for those popular zero-shot conditional samplers.

In this paper, we address two key theoretical research gaps in zero-shot score-mismatched diffusion models: (i) We provide performance guarantees for general score-mismatched diffusion models, extending their applicability beyond the primary focus of existing theoretical studies on the special case of conditional image generation. (ii) We analyze zero-shot conditional diffusion models, which are generally applicable to existing zero-shot samplers such as CCDF (Chung et al., 2022b) and DDNM (Wang et al., 2023) for which the analysis in Xu & Chi (2024) is not applicable (as we discuss above).

1.1 Our Contributions

Technically, the main challenge due to mismatched scores is to analyze the expected tilting factor (Liang et al., 2024) under a mean-perturbed Gaussian, providing an upper bound of the asymptotic orders of all Gaussian non-centralized moments. Our detailed contributions are as follows.

Convergence of General Score-Mismatched DDPM: We provide the *first* non-asymptotic convergence bound on the KL divergence between the target and generated distributions when there is mismatch between the sampling and target scores in DDPM samplers, for general target distributions having finite second moments. We show that the score mismatch at each diffusion step introduces an asymptotic distributional bias that is proportional to the accumulated mismatch. We also provide the first explicit dimensional dependency when the sixth moment of the target distribution exists. Our result is applicable to general forms of mismatch between the target and training scores, which greatly extend the focus of the existing theoretical research on conditional score-mismatch diffusion models.

We then apply our results to zero-shot conditional DDPM samplers, as long as the conditioning involves certain deterministic or random transformations of the data. This provides the first theoretical guarantees for several existing zero-shot samplers, such as CCDF (Chung et al., 2022b) and DDNM (Wang et al., 2023). Notably, the theory in Xu & Chi (2024) does not apply to these samplers, as their analysis relies on the properties of the Markov transition kernel specific to their plug-and-play model. In contrast, our approach is based on the tilting-factor analysis from Liang et al. (2024), which is applicable to general score-mismatched DDPM models. Moreover, the theory in Xu & Chi (2024) is limited to cases where the measurement log-likelihood function is differentiable and bounded and does not provide explicit dependencies on the data dimension. In contrast, our results do not require

[1] Note, however, that some zero-shot methods, such as DPS (Chung et al., 2023) and ΠGDM (Song et al., 2023a), might induce additional computational costs during sampling.

the measurement log-likelihood function to be differentiable or bounded and explicitly characterize the dependencies on the data dimension.

Novel Design of Bias-Optimal Zero-shot Sampler BO-DDNM: Inspired by our convergence analysis of score-mismatched DDPM, we design a novel zero-shot conditional sampler, called the BO-DDNM sampler, which minimizes the asymptotic bias for linear conditional models. Such a sampler coincides with the regular DDNM sampler (Wang et al., 2023) when there is no presence of measurement noise, and achieves faster convergence than both the DDNM and DDNM$^+$ samplers under measurement noise, as shown by our theory and numerical simulations.

Theory for BO-DDNM with Explicit Parameter Dependencies: We provide the convergence bound for the proposed BO-DDNM sampler with explicit dependencies on the dimension d as well as the conditional information y, for various interesting classes of target distributions including those having bounded support and Gaussian mixture. For the case of Gaussian mixture, we further show that three factors positively affect the asymptotic bias: (1) the variance of the measurement noise, (2) the averaged distance between y and the mean of each Gaussian component, and (3) the corresponding correlation coefficient for each component.

1.2 Related Work

We provide a summary of works addressing *unconditional* and *score-matched* diffusion models in Appendix B. Below we discuss related works on *conditional* diffusion models which are closely related to our study here.

Extra-Guided Training: In order to achieve conditional sampling using DDPM models in practice, one method is to introduce conditional guided training, where one either uses an existing classifier (a.k.a., classifier guidance) (Dhariwal & Nichol, 2021) or jointly trains the unconditional and conditional scores (a.k.a., classifier-free guidance (CFG)) (Ho & Salimans, 2022). Here a guidance term is obtained to "guide" the diffusion sampling process at each step such that the sampling scores correspond to the true conditional scores. On the theory side, Wu et al. (2024) investigates the effect of the guidance strength in CFG on Gaussian mixtures, Bradley & Nakkiran (2024) shows that CFG is an instance of predictor-corrector methods, and Chidambaram et al. (2024) finds that CFG might fail to sample correctly on certain mixture targets. There are other theoretical works that investigate sample complexity bounds for conditional score matching for a variety of target distribution models, including the conditional Ising models (Mei & Wu, 2023), those supported on a low-dimensional linear subspace (Yuan et al., 2023), and Hölder smooth and sub-Gaussian conditional models (Fu et al., 2024). Other than stochastic samplers, a conditional ODE sampler is proposed and studied in Chang et al. (2024), which also requires extra training of the conditional score function.

Zero-shot Samplers: To achieve conditional DDPM sampling, a popular method is to use zero-shot conditional samplers, with which one generates a conditional sample using approximated scores. These scores are calculated from the unconditional score estimates and the conditional information using simple (usually linear) functions *without extra-training* (Choi et al., 2021; Chung et al., 2022a;b; 2023; Wang et al., 2023; Song et al., 2023a; Fei et al., 2023). The only theoretical works on the performance of zero-shot DDPM conditional samplers are Xu & Chi (2024); Gupta et al. (2024). In Xu & Chi (2024), a diffusion plug-and-play sampler is proposed which alternates between a diffusion sampling step and a consistency sampling step. The difference of our results from those in Xu & Chi (2024) has been thoroughly discussed in Section 1.1. From an alternative perspective, Gupta et al. (2024) shows that the sampling complexity with zero-shot samplers can take super-polynomial time for some worst-case distribution (among the set of distributions where smooth scores can be efficiently estimated). In contract, our result shows a consistent fact that there exists a non-vanishing asymptotic distributional bias within polynomial time.

2 Problem Setup

In this section, we first provide some background on the (typical) score-*matched* DDPMs. Then, we introduce the score-*mismatched* DDPM samplers and, as a special example, the *conditional* sampling problem and zero-shot samplers.

2.1 Background of Score-Matched DDPMs

The goal of the (typical) score-matched sampling problem is to generate a sample whose distribution is close to the data distribution. To this end, the DDPM algorithm (Ho et al., 2020) is widely used, which consists of a forward process and a reverse process of latent variables. Let $x_0 \in \mathbb{R}^d$ be the initial data, and let $x_t \in \mathbb{R}^d, \forall 1 \leq t \leq T$ be the latent variables. Let Q_0 be the data distribution, and let Q_t (resp., $Q_{t,t-1}$) be the marginal (resp., joint) latent distribution for all $1 \leq t \leq T$.

Forward Process: In the forward process, white Gaussian noise is gradually added to the data: $x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}w_t$, $\forall 1 \leq t \leq T$, where $w_t \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$. Equivalently, this can be expressed as:

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

which means that under Q, the Markov chain $X_0 \to X_1 \to \cdots \to X_T$ holds. Define $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{i=1}^t \alpha_i$ for all $1 \leq t \leq T$. An immediate result by accumulating the steps is that $Q_{t|0}(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I_d)$, or, written equivalently, $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\bar{w}_t$, $\forall 1 \leq t \leq T$, where $\bar{w}_t \sim \mathcal{N}(0, I_d)$ denotes the *aggregated* noise at time t and is independent of x_0. Finally, since each $\bar{w}_t$ is Gaussian, each Q_t $(t \geq 1)$ is absolutely continuous w.r.t. the Lebesgue measure. Let the p.d.f. of each Q_t be q_t, and $q_{t,t-1}$, $q_{t|t-1}$, and $q_{t-1|t}$ for $t \geq 1$ can be similarly defined.

Reverse Process: In the reverse process, the latent variable at time T is first drawn from a standard Gaussian distribution: $x_T \sim \mathcal{N}(0, I_d) =: P_T$. Then, each forward step is approximated by a reverse sampling step. At each time $t = T, T-1, \ldots, 1$, define the *true* reverse process as $x_{t-1} = \mu_t(x_t) + \sigma_t z_t$, where $z \sim \mathcal{N}(0, I_d)$. Here $\sigma_t^2 := \frac{1-\alpha_t}{\alpha_t}$. For the typical DDPM sampling process, $\mu_t(x_t) = \frac{1}{\sqrt{\alpha_t}}(x_t + (1-\alpha_t)\nabla \log q_t(x_t))$. Equivalently, $P_{t-1|t} = \mathcal{N}(x_{t-1}; \mu_t(x_t), \sigma_t^2 I_d)$. Here $\nabla \log q_t(x)$ is called the *score* of q_t, and $\mu_t(x_t)$ is a function of the score. Let P_t be the marginal distributions of x_t in the true reverse process, and let p_t be its corresponding p.d.f. w.r.t. the Lebesgue measure. Define $p_{t-1|t}$ and $p_{t|t-1}$ in a way similar to the forward process.

In practice, one does not have access to $\nabla \log q_t(x_t)$ and thus $\mu_t(x_t)$. Instead, an estimate of $\nabla \log q_t(x_t)$, denoted as $s_t(x_t)$, is used, which results in an estimated $\widehat{\mu}_t(x_t)$ and the *estimated* reverse process: $x_{t-1} = \widehat{\mu}_t(x_t) + \sigma_t z$. Let $\widehat{P}_t$ be the marginal distributions of x_t in the estimated reverse process with the corresponding p.d.f. $\widehat{p}_t$. Note that $\widehat{P}_{t-1|t} = \mathcal{N}(x_{t-1}; \widehat{\mu}_t(x_t), \sigma_t^2 I_d)$ and $\widehat{P}_T = P_T$. Hence, under P and $\widehat{P}$, $X_T \to X_{T-1} \to \cdots \to X_0$ holds.

Performance Metrics: In the case where Q_0 is absolutely continuous w.r.t. the Lebesgue measure, we are interested in measuring the sampling performance through the KL divergence between Q_0 and $\widehat{P}_0$, defined as

$$\mathrm{KL}(Q\|P) := \int \log \tfrac{\mathrm{d}Q}{\mathrm{d}P}\mathrm{d}Q = \mathbb{E}_{X\sim Q}\left[\log \tfrac{q(X)}{p(X)}\right] \geq 0.$$

Indeed, from Pinsker's inequality, the total-variation (TV) distance can be upper bounded as $\mathrm{TV}(Q_0, \widehat{P}_0)^2 \leq \frac{1}{2}\mathrm{KL}(Q_0\|\widehat{P}_0)$. When q_0 does not exist, we use the Wasserstein-2 distance to measure the one-step perturbed performance, which is defined as

$$\mathrm{W}_2(Q, P) := \left\{\min_{\Gamma\in\Pi(Q,P)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2 \, \mathrm{d}\Gamma(x,y)\right\}^{1/2},$$

where $\Pi(Q, P)$ is the set of all joint probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginal distributions Q and P, respectively. Both metrics are widely adopted (Chen et al., 2023a; Benton et al., 2024a).

2.2 Score-Mismatched DDPMs

Differently from the score-*matched* sampling problem, the goal of the score-*mismatched* problem is to sample from a **different** target distribution from the training distribution with which we estimate the scores. Thus, there will be a mismatch between the target score and the estimated score at each diffusion step. Let Q_t $(t \geq 0)$ be the *training* distributions used for training the score. Let $\tilde{Q}_0$ be the *target* distribution that one hopes to generate samples from, and let $\tilde{Q}_t$ $(t \geq 1)$ be its Gaussian-perturbed distributions according to the forward process in (1). Define the posterior mean under the target distributions as $m_t(x_t) := \mathbb{E}_{X_{t-1}\sim\tilde{Q}_{t-1|t}}[X_{t-1}|x_t]$. Note that by Tweedie's formula (Efron, 2011), $m_t(x_t) = \frac{1}{\sqrt{\alpha_t}}(x_t + (1-\alpha_t)\nabla \log \tilde{q}_t(x_t))$. Recall that P_t and $\widehat{P}_t$ are the *sampling* distributions of the true and estimated reverse process, respectively. For general score-mismatched DDPMs, we leave the generic definition of $\mu_t(x_t)$ without providing any particular expression. An example of $\mu_t(x_t)$ is given later in (6), yet our general analysis does not require any particular form for μ_t. With these notations, the *score mismatch* at each step $t \geq 1$ can be defined as

$$\Delta_t(x_t) := \frac{\sqrt{\alpha_t}}{1-\alpha_t}\left(\mathbb{E}_{X_{t-1}\sim\tilde{Q}_{t-1|t}}[X_{t-1}|x_t] - \mathbb{E}_{X_{t-1}\sim P_{t-1|t}}[X_{t-1}|x_t]\right) = \frac{\sqrt{\alpha_t}}{1-\alpha_t}(m_t(x_t) - \mu_t(x_t)). \tag{2}$$

The goal, then, is to provide an upper bound on the distributional dissimilarity between the target distribution $\tilde{Q}_0$ and the sampling distribution $\widehat{P}_0$. We use the same metrics as those defined in Section 2.1 to evaluate the performance of the score-mismatched DDPM.

2.3 Zero-Shot Conditional DDPMs

One interesting example of score mismatch is the zero-shot conditional sampling problem. Differently from the unconditional counterpart, the conditional sampling problem aims to obtain a sample that aligns in particular with the provided conditioning. Define $y \in \mathbb{R}^p$ to be the conditioned information about x_0. Specifically, let $y = h(x_0)$, where $h(\cdot)$ is some arbitrary (deterministic or random) function of only x_0 (apart from independent noise). Note that general score-mismatched DDPMs can be specialized to zero-shot conditional samplers with the following notations:

$$\tilde{Q}_t = Q_{t|y}, \quad m_t = m_{t,y}, \quad \mu_t = \mu_{t,y}, \quad \text{and } \Delta_t = \Delta_{t,y}. \tag{3}$$

Linear Conditional Models: In practice, one commonly adopted model is the linear conditional model (Jalal et al., 2021; Wang et al., 2023; Song et al., 2023a), defined as

$$y := Hx_0 + n, \tag{4}$$

where $H \in \mathbb{R}^{p\times d}$ ($p \leq d$) is a deterministic matrix and $n \sim \mathcal{N}(0, \sigma_y^2 I_p)$ is the measurement noise, which is assumed to be Gaussian and independent of x_0. For the case where there is no measurement noise, let $\sigma_y^2 = 0$ and thus $n = 0$ almost surely. In applications like image super-resolution and inpainting (Wang et al., 2023), H admits a simple form of a 0-1 diagonal matrix, where the 1's occur only on the diagonal and at those locations corresponding to the provided pixels. In these scenarios, both H and y are fixed and given. The linear conditional model is studied in Section 4.

Conditional Forward Process for Linear Models: Write the Moore–Penrose pseudo-inverse of H as $H^\dagger$, and note that $H^\dagger H$ is an orthogonal projection matrix. With this notation, under (4), we can re-express the forward process in (1) as

$$x_t = \sqrt{\bar{\alpha}_t}(I_d - H^\dagger H)x_0 + \sqrt{\bar{\alpha}_t}H^\dagger y - \sqrt{\bar{\alpha}_t}H^\dagger n + \sqrt{1-\bar{\alpha}_t}\bar{w}_t.$$

Here, since n is independent of $\bar{w}_t$, for fixed x_0 and y, we have that, for all $t \geq 1$,

$$Q_{t|0,y}(x_t|x_0, y) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}(I_d - H^\dagger H)x_0 + \sqrt{\bar{\alpha}_t}H^\dagger y, \bar{\alpha}_t\sigma_y^2 H^\dagger (H^\dagger)^\intercal + (1-\bar{\alpha}_t)I_d). \tag{5}$$

Also, since the forward process is a Markov chain, we have that $Q_{t|t-1,y} = Q_{t|t-1}$ for all $t \geq 1$.

Zero-shot Conditional Sampler for Linear Models: We employ the *zero-shot* conditional sampler for linear conditional models in the following form: $x_{t-1} = \mu_{t,y}(x_t) + \sigma_t z_t$, where

$$\mu_{t,y}(x_t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t + (1-\alpha_t)g_{t,y}(x_t)\right), \quad g_{t,y} := (I_d - H^\dagger H)\nabla \log q_t(x_t) + f_{t,y}(x_t). \tag{6}$$

Here $f_{t,y}(x_t)$ is a simple function of y and x_t computable *without extra training* and such that $(I_d - H^\dagger H)f_{t,y}(x) \equiv 0$ for all $x \in \mathbb{R}^d$. Intuitively, $f_{t,y}$ characterizes the score rectification in the range space of $H^\dagger H$. Indeed, many zero-shot samplers in the literature have such $f_{t,y}(x_t)$'s that satisfy (6) (see Appendix D). Now, with the linear model in (4) and the zero-shot conditional sampler in (6), the score mismatch at each time $t \geq 1$ is equal to

$$\Delta_{t,y}(x_t) = (I_d - H^\dagger H)(\nabla \log q_{t|y}(x_t) - \nabla \log q_t(x_t)) + (H^\dagger H)\nabla \log q_{t|y}(x_t) - f_{t,y}(x_t). \tag{7}$$

3 DDPM under General Score Mismatch

In this section, we provide convergence guarantees for general score-mismatched DDPM samplers under a general target distribution $\tilde{Q}_0$. Throughout this section we keep the generic definition for score mismatch Δ_t as in (2), without assuming any particular expression for μ_t.

3.1 Technical Assumptions

We will analyze general score-mismatched DDPMs under the following technical assumptions.

Assumption 1 (Finite Second Moment)**.** There exists a constant $M_2 < \infty$ (that does not depend on d and T) such that $\mathbb{E}_{X_0 \sim \tilde{Q}_0} \|X_0\|^2 \leq dM_2$.

The first Assumption 1 is commonly adopted in the analyses of score-matched DDPM samplers (Chen et al., 2023a;d; Liang et al., 2024).

Assumption 2 (Posterior Mean Estimation)**.** The estimated posterior mean $\widehat{\mu}_t$ at $t = 1, \dots, T$ satisfy

$$\frac{1}{T} \sum_{t=1}^{T} \frac{\alpha_t}{(1-\alpha_t)^2} \mathbb{E}_{X_t \sim \tilde{Q}_t} \|\widehat{\mu}_t(X_t) - \mu_t(X_t)\|^2 \leq \varepsilon^2, \text{ where } \varepsilon^2 = \tilde{O}(T^{-2}).$$

The above Assumption 2 is made for the score estimation error for the general mismatched setting, where we leave generic definitions of μ_t and $\widehat{\mu}_t$. For zero-shot conditional samplers in linear models, this assumption is weaker than that for the estimation error for unconditional scores (see (9)). Compared with the score-matched case, the estimation error needs to be achieved at a higher accuracy because of the extra error term when there is score mismatch (Lemma 2). Such a higher level of estimation accuracy also occurs in previous theoretical studies for accelerated DDPM samplers (Li et al., 2024a, Theorem 3.2).

Assumption 3 (Regular Derivatives)**.** For all $t \geq 1$ where $\tilde{q}_{t-1}$ exists, $\ell \geq 1$, and $\boldsymbol{a} \in [d]^p$ where $|\boldsymbol{a}| = p \geq 1$,

$$\mathbb{E}_{X_t \sim \tilde{Q}_t} |\partial_{\boldsymbol{a}}^p \log \tilde{q}_t(X_t)|^\ell = O\left(1\right), \quad \mathbb{E}_{X_t \sim \tilde{Q}_t} |\partial_{\boldsymbol{a}}^p \log \tilde{q}_{t-1}(m_t(X_t))|^\ell = O\left(1\right).$$

The above Assumption 3 is useful for our tilting-factor based analysis, which guarantees that all (higher-order) Taylor polynomials of $\log \tilde{q}_t$ are well controlled in expectation. It is rather soft, and it can be verified when $\tilde{Q}_0$ has finite variance (under early-stopping) (Liang et al., 2024).

Assumption 4 (Bounded Mismatch)**.** For all $t \geq 1$ where $\tilde{q}_{t-1}$ exists, and $\ell \geq 2$,

$$\mathbb{E}_{X_t \sim \tilde{Q}_t} \|\Delta_t(X_t)\|^\ell = O(\bar{\alpha}_t).$$

The above Assumption 4 is used to characterize the amount of mismatch at each time $t \geq 1$. The $\bar{\alpha}_t := \prod_{i=1}^{t} \alpha_i$ is necessary for the overall bias to be bounded.

In the paper, Assumptions 3 and 4 have been established in two cases of zero-shot conditional sampling: (i) where Q_0 has bounded support for any H, using a special α_t in (8) (see the proof of Theorem 4); and (ii) where Q_0 is Gaussian mixture and $H = (I_p \quad 0)$ (see Lemma 8). For Case (i), the assumption that Q_0 has bounded support has wide applicability in practice (e.g., images (Ho et al., 2020; Wang et al., 2023)) and is commonly made in many theoretical investigations of the score-matched DDPM (Li et al., 2024a;c).

Finally, note that when $\tilde{q}_0$ does not exist (e.g., for images (Ho et al., 2020; Wang et al., 2023)), Assumptions 3 and 4 are required only for $t \geq 2$.

3.2 Convergence Bound

To present the main result, we first define a set of noise schedule as follows.

Definition 1 (Noise Schedule)**.** For all sufficiently large T, set the step size α_t's to satisfy

$$1 - \alpha_t \lesssim \frac{\log T}{T}, \ \forall 1 \leq t \leq T, \quad \text{and} \quad \bar{\alpha}_T := \prod_{t=1}^{T} \alpha_t = o\left(\frac{1}{T}\right).$$

An example of α_t that satisfies Definition 1 is $1 - \alpha_t \equiv \frac{c \log T}{T}$, $\forall t \geq 1$ with $c > 1$. Then, $\bar{\alpha}_T = \left(1 - \frac{c \log T}{T}\right)^T = \exp\left(T \log\left(1 - \frac{c \log T}{T}\right)\right) = O\left(e^{T \frac{-c \log T}{T}}\right) = o\left(T^{-1}\right)$.

The following Theorem 1 provides an upper bound on the KL-divergence between the target distribution $\tilde{Q}_0$ and the sampling distribution $\widehat{P}_0$, as a function of (general) score-mismatch Δ_t at each time $t \geq 1$. Theorem 1 is the *first* convergence result for score-mismatched DDPM samplers for any smooth $\tilde{Q}_0$ that has finite second moment (along with some mild regularity conditions).

Theorem 1 (DDPM under Score Mismatch)**.** *Suppose that $\tilde{Q}_0$ has a p.d.f. $\tilde{q}_0$ which is analytic, and suppose that Assumptions 1 to 4 are satisfied. Then, with the α_t chosen to satisfy Definition 1, the distribution $\widehat{P}_0$ from the score-mismatched DDPM satisfies*

$$\mathrm{KL}(\tilde{Q}_0 \| \widehat{P}_0) \lesssim \mathcal{W}_{oracle} + \mathcal{W}_{bias} + \mathcal{W}_{vanish}, \quad \textit{where}$$

$$\begin{aligned}
\mathcal{W}_{oracle} &= \textstyle\sum_{t=1}^T \frac{(1-\alpha_t)^2}{2\alpha_t} \mathbb{E}_{X_t \sim \tilde{Q}_t} \left[\mathrm{Tr}\Big(\nabla^2 \log \tilde{q}_{t-1}(m_t(X_t)) \nabla^2 \log \tilde{q}_t(X_t) \Big) \right] + (\log T)\varepsilon^2 \\
\mathcal{W}_{bias} &= \textstyle\sum_{t=1}^T (1-\alpha_t) \mathbb{E}_{X_t \sim \tilde{Q}_t} \left\| \Delta_t(X_t) \right\|^2 \\
\mathcal{W}_{vanish} &= \textstyle\sum_{t=1}^T \frac{1-\alpha_t}{\sqrt{\alpha_t}} \mathbb{E}_{X_t \sim \tilde{Q}_t} \Big[(\nabla \log \tilde{q}_{t-1}(m_t(X_t)) - \sqrt{\alpha_t} \nabla \log \tilde{q}_t(X_t))^\intercal \Delta_t(X_t) \Big] \\
&\quad - \textstyle\sum_{t=1}^T \frac{(1-\alpha_t)^2}{2\alpha_t} \mathbb{E}_{X_t \sim \tilde{Q}_t} \left[\Delta_t(X_t)^\intercal \nabla^2 \log \tilde{q}_{t-1}(m_t(X_t)) \Delta_t(X_t) \right] \\
&\quad + \textstyle\sum_{t=1}^T \frac{(1-\alpha_t)^2}{3! {\alpha_t}^{3/2}} \mathbb{E}_{X_t \sim \tilde{Q}_t} \Big[3 \sum_{i=1}^d \partial^3_{iii} \log \tilde{q}_{t-1}(m_t(X_t)) \Delta_t(X_t)^i \\
&\qquad + \textstyle\sum_{\substack{i,j=1 \\ i \neq j}}^d \partial^3_{iij} \log \tilde{q}_{t-1}(m_t(X_t)) \Delta_t(X_t)^j \Big] + \max_{t \geq 1} \sqrt{\mathbb{E}_{X_t \sim \tilde{Q}_t} \left\| \Delta_t(X_t) \right\|^2} (\log T) \varepsilon.
\end{aligned}$$

When $\tilde{Q}_0$ does not have a p.d.f., a similar upper bound is applied to $\mathrm{KL}(\tilde{Q}_1 \| \widehat{P}_1)$ such that $\mathrm{W}_2(\tilde{Q}_1, \tilde{Q}_0)^2 \lesssim (1-\alpha_1)d$ (see Corollary 1 in Appendix F.5).

To explain the three error terms in Theorem 1, $\mathcal{W}_{\text{oracle}}$ captures the error assuming that one has access to (a close estimate of) $\nabla \log \tilde{q}_t, \ \forall t \geq 1$. This error is independent of the score mismatch Δ_t, and it decays as $\tilde{O}(T^{-1})$ under Assumption 3 (Liang et al., 2024, Theorem 1). The remaining two error terms $\mathcal{W}_{\text{bias}}$ and $\mathcal{W}_{\text{vanish}}$ arise from the mismatched sampling process. Both terms become zero if $\Delta_t \equiv 0$ for all $t \geq 1$, which corresponds to the score-matched case. Under Assumptions 3 and 4, $\mathcal{W}_{\text{vanish}}$ decays as $\tilde{O}(T^{-1})$ under an additional mild condition (see Lemma 5 in Appendix G), and $\mathcal{W}_{\text{bias}}$ asymptotically approaches a constant. Combining all three terms, score mismatch causes an asymptotic distributional bias between $\tilde{Q}_0$ and $\widehat{P}_0$.

To further understand $\mathcal{W}_{\text{bias}}$, note that $1-\alpha_t$ is usually summable under Assumption 4 (cf. Lemmas 7 and 10). Thus, the bias can be further upper-bounded by the maximum step-wise mismatch $\max_{t \geq 1} \mathbb{E}_{X_t \sim \tilde{Q}_t} \left\| \Delta_t(X_t) \right\|^2$. In case that $\mu_t(x_t) = \frac{1}{\sqrt{\alpha_t}} (x_t + (1-\alpha_t) g_t(x_t))$ (as for the zero-shot sampler in (6)), define a measure $\tilde{P}_t$ such that $g_t(x_t) = \nabla \log \tilde{p}_t(x_t)$. Then, from (2),

$$\mathbb{E}_{X_t \sim \tilde{Q}_t} \left\| \Delta_t(X_t) \right\|^2 = \mathbb{E}_{X_t \sim \tilde{Q}_t} \left\| \nabla \log \frac{\tilde{q}_t(X_t)}{\tilde{p}_t(X_t)} \right\|^2 =: \mathcal{F}(\tilde{Q}_t \| \tilde{P}_t).$$

where $\mathcal{F}(Q \| P)$ denotes the *Fisher divergence* (or called relative Fisher information) between Q and P. In Section 4, this distributional bias $\mathcal{W}_{\text{bias}}$ inspires us to design a novel zero-shot DDPM sampler, the BO-DDNM sampler, that minimizes the asymptotic bias.

Next we provide an upper bound with explicit dimensional dependency, for any Q_0 that has finite sixth moment such as Gaussian mixture Q_0's and those Q_0's having bounded support. To this end, we consider a special noise schedule first proposed in Li et al. (2024c):

$$1-\alpha_1 = \delta, \quad 1-\alpha_t = \frac{c \log T}{T} \min \left\{ \delta \left(1 + \frac{c \log T}{T} \right)^t, 1 \right\}, \ \forall 2 \leq t \leq T \tag{8}$$

for any constants (c, δ) such that $c > 1$ and $\delta e^c > 1$. Note that this noise schedule corresponds to early-stopping in the literature (Chen et al., 2023a; Benton et al., 2024a). With the α_t in (8), we can show that the regularity condition Assumption 3 holds for a quite general set of distributions (see Lemma 5 in Appendix G).

Theorem 2. *Suppose that $\mathbb{E}_{X_0 \sim Q_0} \left\| X_0 \right\|^6 \lesssim d^3$. Further, suppose that Δ_t satisfies that $\mathbb{E}_{X_t \sim \tilde{Q}_t} \left\| \Delta_t(X_t) \right\|^4 \lesssim \frac{\bar{\alpha}_t^2}{(1-\bar{\alpha}_t)^{2r}} d^{2\gamma}$ with some $\gamma, r \geq 1$ for all $t \geq 2$. Then, if the estimation error satisfies Assumption 2 and if Δ_t satisfies Assumption 4, with the α_t in (8) such that $\delta \ll 1$ and*

$c \asymp \log(1/\delta)$, *we have, for some* $\tilde{Q}_1$ *such that* $\mathrm{W}_2(\tilde{Q}_1, \tilde{Q}_0)^2 \lesssim \delta d$,

$$\mathrm{KL}(\tilde{Q}_1 \| \widehat{P}_1) \lesssim d^{\gamma}\delta^{-r}\left(1 - \frac{2\log(1/\delta)\log T}{T}\right) + \max\{d^{(3+\gamma)/2}\delta^{-\frac{r+2}{2}}, d^{1+\gamma}\delta^{-(r-1)}\}\frac{(\log T)^2}{T} + d^{\gamma/2}\delta^{-r/2}(\log T)\varepsilon.$$

Note that Theorem 2 provides the *first* performance guarantee with *explicit* dimensional dependence for general score-mismatched DDPMs. Here the finite sixth moment is a technical condition to guarantee small expected difference of the first-order Taylor polynomial in case of mismatched scores (see Lemma 5 in Appendix G). Later, Theorem 2 will be useful to provide guarantees for zero-shot conditional samplers under linear models (Theorem 4).

4 Zero-shot Conditional DDPM Samplers

As we discuss before, an important scenario of score-mismatched diffusion models is the zero-shot conditional problem, where certain information y is given. In this section, we apply our general results for score-mismatch DDPMs in Section 3 to studying zero-shot conditional DDPM samplers, with the notations in (3) for $y = h(x_0)$ for some arbitrary (deterministic or random) $h(\cdot)$.

In the following, we are particularly interested in the linear conditional model in (4). We take the same Assumptions 1, 3 and 4 (albeit with changed notations), and further adopt the following common assumption on the *unconditional* score estimation (Chen et al., 2023a;d; Liang et al., 2024).

Assumption 5 (Estimation Error of Unconditional Score)**.** Suppose that s_t satisfies

$$\tfrac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{X_t\sim Q_{t|y}}\left\|s_t(X_t) - \nabla\log q_t(X_t)\right\|^2 \le \varepsilon^2, \text{ where } \varepsilon^2 = \tilde{O}(T^{-2}).$$

Note that, with the zero-shot sampler defined in (6), since $\left\|I_d - H^\dagger H\right\| = 1$, we have, $\forall x \in \mathbb{R}^d$,

$$\left\|\widehat{\mu}_{t,y} - \mu_{t,y}\right\|^2 = \frac{(1-\alpha_t)^2}{\alpha_t}\left\|(I_d - H^\dagger H)(s_t - \nabla\log q_t)\right\|^2 \le \frac{(1-\alpha_t)^2}{\alpha_t}\left\|s_t - \nabla\log q_t\right\|^2. \quad (9)$$

Therefore, Assumption 5 directly implies Assumption 2, and thus Theorem 1 (as well as Corollary 1) still holds under Assumptions 1 and 3 to 5.

4.1 A Novel Bias-Optimal Zero-shot Sampler

Guided by the performance guarantee characterized in Theorem 1, we will propose a novel *optimized* zero-shot condition sampler. With the zero-shot sampler defined in (6), the goal is to choose the $f_{t,y}$ function that minimizes the convergence error for each $y \in \mathbb{R}^p$ and $t \ge 1$.

Specifically, it is observed in Theorem 1 that the convergence error in terms of the KL-divergence will have an asymptotic distributional bias given by $\mathcal{W}_{\text{bias}}$. As follows, we characterize an optimal $f_{t,y}$ that minimizes $\mathcal{W}_{\text{bias}}$, which thus yields a corresponding bias-optimal zero-shot sampler.

Theorem 3. *Define* $\Sigma_{t|0,y} := \bar{\alpha}_t\sigma_y^2 H^\dagger(H^\dagger)^\intercal + (1-\bar{\alpha}_t)I_d$. *For any* Q_0 *and* $t \ge 1$, *we have*

$$\nabla\log q_{t|y}(x_t) = \Sigma_{t|0,y}^{-1}(\sqrt{\bar{\alpha}_t}H^\dagger y - x_t) + \tfrac{\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t}(I_d - H^\dagger H)\mathbb{E}_{Q_{0|t,y}}[X_0|x_t, y].$$

Also, recall the sampler in (6) *and define* $f^*_{t,y}$ *as*

$$f^*_{t,y}(x_t) := \Sigma_{t|0,y}^{-1}\left(\sqrt{\bar{\alpha}_t}H^\dagger y - H^\dagger H x_t\right). \quad (10)$$

Also recall $\Delta_{t,y}$ *from* (7)*. Then,* $f^*_{t,y}$ *satisfies that, for all* $t \ge 1$ *and fixed* $y \in \mathbb{R}^p$,

$$f^*_{t,y} \in \mathop{\arg\min}_{f_{t,y}:(I_d - H^\dagger H)f_{t,y}\equiv 0}\left\|\Delta_{t,y}\right\|^2, \quad Q_{t|y}\text{–almost surely.}$$

The sampler $f^*_{t,y}(x_t)$ defined in (10) provides a bias-optimal zero-shot conditional DDPM sampler. In the case with $\sigma_y = 0$, such an optimal sampler coincides with the regular DDNM sampler in Wang et al. (2023) (see Appendix D). Thus, we call this sampler as **Bias-Optimal (BO) DDNM** sampler. With (10), we can also calculate the minimum step-wise mismatch as

$$\min_{f_{t,y}:(I_d - H^\dagger H)f_{t,y}\equiv 0}\mathbb{E}_{X_t\sim Q_{t|y}}\left\|\Delta_{t,y}\right\|^2 = \mathbb{E}_{X_t\sim Q_{t|y}}\left\|\nabla\log\tfrac{q_{t|y}(X_t)}{q_t(X_t)}\right\|^2_{(I_d - H^\dagger H)},$$

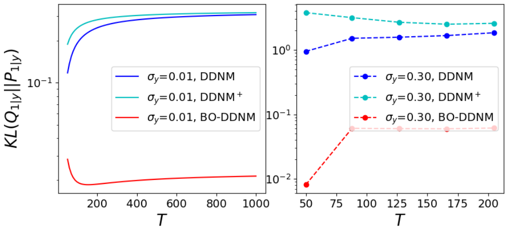

Figure 1: Comparison of BO-DDNM, DDNM and DDNM$^+$ for Gaussian (left) and Gaussian mixture (right) Q_0 under measurement noise.

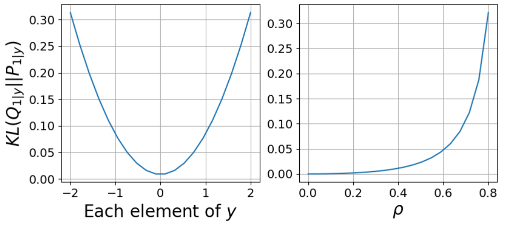

Figure 2: Distributional bias as a function of the conditioning y (left) and the correlation coefficient ρ (right) for Gaussian Q_0.

which is the projected Fisher divergence between $Q_{t|y}$ and Q_t on $\mathrm{range}(I_d - H^\dagger H)$.

In the following lemma, we provide the performance bound for BO-DDNM when Q_0 has bounded support. For comparison, we also provide the theoretical performance of vanilla DDNM, denoted as $f^N_{t,y}$.

Theorem 4 (BO-DDNM vs. DDNM)**.** *Suppose that $\|X_0\|^2 \leq R^2 d$ a.s. under Q_0. Suppose that Assumptions 1 and 5 hold. Then, with the conditional sampler $f^*_{t,y}$ in* (10)*, Theorem 2 holds with $\gamma = 1$ and $r = 2$. Also, with the conditional sampler $f^N_{t,y} := (1-\bar{\alpha}_t)^{-1}\left(\sqrt{\bar{\alpha}_t} H^\dagger y - H^\dagger H x_t\right)$, if further $\left\|H^\dagger\right\| \lesssim 1$, then Theorem 2 holds with $\gamma = 1$ and $r = 4$.*

Theorem 4 establishes the *first* result applicable to DDNM-type zero-shot conditional samplers for any linear conditional models on those target distributions having bounded support.

Advantage of BO-DDNM over DDNM and DDNM$^+$: When there is positive measurement noise, Theorem 4 indicates that our BO-DDNM sampler that uses $f_{t,y} = f^*_{t,y}$ enjoys a smaller asymptotic bias than DDNM that uses $f^N_{t,y}$ with the α_t in (8) (δ^{-2} vs. δ^{-4}). Note that the DDNM sampler corresponds to $f_{t,y} = f^N_{t,y}$ (see Appendix D). Such an advantage is also demonstrated by our numerical experiment. In Figure 1, we numerically compared modified conditional zero-shot sampler (as given in (10)) with the DDNM and DDNM$^+$ sampler for both Gaussian and Gaussian mixture Q_0's at different levels of measurement noise. It is observed that the optimal BO-DDNM sampler achieves a much lower bias than both the DDNM and the DDNM$^+$ samplers numerically, especially when σ_y^2 becomes large.

4.2 BO-DDNM Sampler for Gaussian Mixture Q_0

In this section, we focus on the convergence dependency on other system parameters of the BO-DDNM sampler, including the chosen y. In particular, we restrict our attention to Gaussian mixture Q_0's and to a special conditional model, where $H = (I_p \quad 0)$. This choice can be seen in many applications, such as image super-resolution and inpainting (after reorganizing the pixels), where I_p corresponds to the locations of the given pixels (Wang et al., 2023; Song et al., 2023a). We assume positive measurement noise. We introduce the notation $[\Sigma_0]_{ab}$ to denote the variance components that correspond to the space of $a \times b$ where $a, b \in \{y, \bar{y}\}$.

The following Proposition 1 gives an upper bound on the asymptotic bias for Gaussian mixture Q_0.

Proposition 1. *Suppose that Q_0 is Gaussian mixture with equal variance, 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 $\mathcal{N}(\mu_{0,n}, \Sigma_0)$ and $\pi_n \in [0,1]$ is the mixing coefficient with $\sum_{n=1}^N \pi_n = 1$. Suppose that $H = (I_p \quad 0)$, and adopt $f^*_{t,y}$ in* (10) *and α_t in Definition 1. Write $\lambda_1 \geq \cdots \geq \lambda_d > 0$ and $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_{d-p} > 0$ as the eigenvalues of Σ_0 and $[\Sigma_0]_{\bar{y}\bar{y}}$, respectively. Then,*

$$\begin{aligned}
&\mathbb{E}_{X_t \sim Q_{t|y}} \left\|\Delta_{t,y}(X_t)\right\|^2 \\
&\lesssim \bar{\alpha}_t d + \bar{\alpha}_t^2 \tfrac{\|[\Sigma_0]_{y\bar{y}}\|^2}{\min\{\tilde{\lambda}_{d-p},1\}^2 \min\{\lambda_d,1\}^2} \max\left\{ d(\lambda_1 + \sigma_y^2) + \textstyle\sum_{n=1}^N \pi_n \left\|H^\dagger y - H^\dagger H \mu_{0,n}\right\|^2, d \right\} \\
&\lesssim \bar{\alpha}_t \left(d + \textstyle\sum_{n=1}^N \pi_n \left\|H^\dagger y - H^\dagger H \mu_{0,n}\right\|^2 \right).
\end{aligned}$$

Proposition 1 indicates that three factors affect (an upper bound on) the asymptotic bias. (i) The measurement noise variance σ_y^2 determined by the system nature has an increasing effect on the bias. (ii) The averaged distance $\sum_{n=1}^N \pi_n \left\|H^\dagger y - H^\dagger H \mu_{0,n}\right\|^2$ between $H^\dagger y$ and $H^\dagger H \mu_{0,n}$ captures the quadratic dependency in y, as illustrated in the left plot of Figure 2. (iii) The correlation between Hx_0 and $(I_d - H^\dagger H)x_0$ of each mixture component contributes positively to the bias, which is contained in the factor $\frac{\|[\Sigma_0]_{y\bar{y}}\|^2}{\min\{\lambda_d, 1\}^2}$. To see this, consider $\sigma_y^2 = 0$ and a specific Gaussian example with $d = 2$, $p = 1$, and $\Sigma_0 = \begin{pmatrix} \sigma_{11}^2 & \rho\sigma_{11}\sigma_{22} \\ \rho\sigma_{11}\sigma_{22} & \sigma_{22}^2 \end{pmatrix}$. As the correlation coefficient ρ increases, Σ_0 becomes closer to be singular, and thus λ_d decreases to 0. Also, $\left\|[\Sigma_0]_{y\bar{y}}\right\|^2 = \rho^2 \sigma_{11}^2 \sigma_{22}^2$ increases quadratically with ρ. Hence, this factor $\frac{\|[\Sigma_0]_{y\bar{y}}\|^2}{\min\{\lambda_d, 1\}^2}$ grows unboundedly as $\rho \to 1$, as does $\mathbb{E}_{X_t \sim Q_{t|y}} \left\|\Delta_{t,y}(X_t)\right\|^2$. Such dependency on the correlation is illustrated numerically in the right plot of Figure 2.

The following theorem characterizes the conditional KL divergence when Q_0 is mixture Gaussian. In particular, we can show Assumption 4 holds with any α_t that satisfies Definition 1 when Q_0 is Gaussian mixture (see Lemma 8 in Appendix I.5).

Theorem 5. *Suppose the same conditions in Proposition 1 hold and $\sigma_y^2 > 0$. Suppose that Assumption 5 holds. Take $f_{t,y}^*$ in* (10) *and α_t that further satisfies $\sum_{t=1}^T (1 - \alpha_t)\bar{\alpha}_t = 1 + o(1)$. Then,*

$$\mathrm{KL}(Q_{0|y} \| \widehat{P}_{0|y}) \lesssim \left(d + \sum_{n=1}^N \pi_n \left\|H^\dagger y - H^\dagger H \mu_{0,n}\right\|^2\right) +$$

$$\left(d^2 + \sum_{n=1}^N \pi_n \left\|H^\dagger y - H^\dagger H \mu_{0,n}\right\|^4\right) \frac{(\log T)^2}{T} + \sqrt{d + \sum_{n=1}^N \pi_n \left\|H^\dagger y - H^\dagger H \mu_{0,n}\right\|^2} (\log T)\varepsilon.$$

Although Proposition 1 and Theorem 5 assume $H = (I_p \quad 0)$, extension to general H is straightforward by modifying the proof of Lemma 8 and using the fact that $\left\|H^\dagger H\right\| = \left\|I_d - H^\dagger H\right\| = 1$.

This is the *first* convergence result for zero-shot samplers where explicit dependency on the conditioning y is derived for Gaussian mixture targets. Note that the extra condition on α_t can be verified for both constant α_t (Lemma 10) and that in (8) (Lemma 7). Among the three terms in Theorem 5, the first term is the asymptotic bias analyzed in Proposition 1. Since the last two terms decrease to zero as $T \to \infty$, the asymptotic KL divergence will also approach some non-zero limit of order d.

The proof of Theorem 5 is non-trivial because from Theorem 1 we need to figure out the dependency on y in all first three orders of partial derivatives of a Gaussian mixture density, which is generally hard to express. To this end, we restrict focus to a particular linear model where explicit dependency can be sought. The result can be extended to the case of $\sigma_y^2 = 0$ with the α_t in (8) (see Remark 2).

5 Conclusion

In this paper, we have provided convergence guarantees for the general score-mismatched diffusion models, which are specialized to zero-shot conditional samplers. For linear conditional models, we also designed an optimal BO-DDNM sampler that minimizes the asymptotic bias, for which we showed the dependencies on the system parameters. One future direction is to explore zero-shot samplers that use higher-order derivatives of the log-densities, which might achieve better convergence results.

References

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.

Arwen Bradley and Preetum Nakkiran. Classifier-free guidance is a predictor-corrector. *arXiv preprint arXiv:2408.09000*, 2024.

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.

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.

Muthu Chidambaram, Khashayar Gatmiry, Sitan Chen, Holden Lee, and Jianfeng Lu. What does guidance do? a fine-grained analysis in a simple setting. *arXiv preprint arXiv:2409.13074*, 2024.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14367–14376, October 2021.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *Advances in Neural Information Processing Systems*, 2022a.

Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster accelerating conditional diffusion models for inverse problems through stochastic contraction. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022b.

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 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.

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, 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.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021.

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

Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9935–9946, June 2023.

John Torjus Flåm. The linear model under gaussian mixture inputs: Selected problems in communications, 2013.

Hengyu Fu, Zhuoran Yang, Mengdi Wang, and Minshuo Chen. Unveil conditional diffusion models with classifier-free guidance: A sharp statistical theory. *arXiv preprint arXiv:2403.11968*, 2024.

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.

Shivam Gupta, Ajil Jalal, Aditya Parulekar, Eric Price, and Zhiyang Xun. Diffusion posterior sampling is computationally intractable. *arXiv preprint arXiv:2402.12727*, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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*, 2024.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G Dimakis, and Jon Tamir. Robust compressed sensing mri with deep generative priors. In *Advances in Neural Information Processing Systems*, 2021.

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.

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

Xin Li, Yulin Ren, Xin Jin, Cuiling Lan, Xingrui Wang, Wenjun Zeng, Xinchao Wang, and Zhibo Chen. Diffusion models for image restoration and enhancement – a comprehensive survey. *arXiv preprint arXiv:2308.09388*, 2023.

Yuchen Liang, Peizhong Ju, Yingbin Liang, and Ness Shroff. Non-asymptotic convergence of discrete-time diffusion models: New approach and improved rate. *arXiv preprint arXiv:2402.13901*, 2024.

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.

Brian B. Moser, Arundhati S. Shanbhag, Federico Raue, Stanislav Frolov, Sebastián M. Palacio, and Andreas Dengel. Diffusion models, image super-resolution and everything: A survey. *arXiv preprint arXiv:2401.00736*, 2024.

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.

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

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

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.

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, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023a.

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

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *The Eleventh International Conference on Learning Representations*, 2023.

Yuchen Wu, Minshuo Chen, Zihao Li, Mengdi Wang, and Yuting Wei. Theoretical insights for diffusion guidance: A case study for gaussian mixture models. *arXiv preprint arXiv:2403.01639*, 2024.

Xingyu Xu and Yuejie Chi. Provably robust score-based diffusion posterior sampling for plug-and-play image reconstruction. *arXiv preprint arXiv:2403.17042*, 2024.

Hui Yuan, Kaixuan Huang, Chengzhuo Ni, Minshuo Chen, and Mengdi Wang. Reward-directed conditional diffusion: Provable distribution estimation and reward improvement. In *Thirty-seventh Conference on Neural Information Processing Systems*, 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.