Theory of Consistency Diffusion Models: Distribution Estimation Meets Fast Sampling

Zehao Dou [1] **Minshuo Chen** [2] **Mengdi Wang** [2] **Zhuoran Yang** [1]

Abstract

Diffusion models have revolutionized various application domains, including computer vision and audio generation. Despite the state-of-the-art performance, diffusion models are known for their slow sample generation due to the extensive number of steps involved. In response, consistency models have been developed to merge multiple steps in the sampling process, thereby significantly boosting the speed of sample generation without compromising quality. This paper contributes towards the first statistical theory for consistency models, formulating their training as a distribution discrepancy minimization problem. Our analysis yields statistical estimation rates based on the Wasserstein distance for consistency models, matching those of vanilla diffusion models. Additionally, our results encompass the training of consistency models through both distillation and isolation methods, demystifying their underlying advantage.

1 Introduction

Diffusion models have reached state-of-the-art performance in cross-domain applications, including computer vision (Song & Ermon, 2019; Dathathri et al., 2019; Ho et al., 2020; Song et al., 2020c), audio generation (Kong et al., 2020; Chen et al., 2020), language generation (Li et al., 2022; Yu et al., 2022; Lovelace et al., 2022), reinforcement learning and control (Pearce et al., 2023; Chi et al., 2023; Hansen-Estruch et al., 2023; Reuss et al., 2023), as well as computational biology (Lee et al., 2022c; Luo et al., 2022; Gruver et al., 2023). These break-through performances are enabled by the unique design in the diffusion models. Specifically, diffusion models utilize the forward and backward processes to generate new samples. In the forward process, a clean data point is progressively contaminated by random noise, while the backward process attempts to remove the noise iteratively (typically taking 500 to 1000 steps (Song & Ermon, 2019)) with the help of a specific type of neural network known as a score neural network.

Due to the enormous size of the score neural network, e.g., the smallest stable diffusion model uses a network of more than 890M parameters (Rombach et al., 2022), the sample generation speed of diffusion models is limited (Song et al., 2023), compared to Generative Adversarial Networks (Goodfellow et al., 2020) and AutoEncoders (Kingma et al., 2019). To overcome this shortcoming, there are extensive methodological studies aiming to accelerate diffusion models. Notable methods include using stride in sampling to reduce the number of backward steps (Nichol & Dhariwal, 2021; Song & Ermon, 2020; Lu et al., 2022), changing the backward process to a deterministic probabilistic flow (Song et al., 2020a; Karras et al., 2022; Zhang et al., 2022), and utilizing pretrained variational autoencoders to reduce the data dimensionality before applying diffusion models (Rombach et al., 2022). These methods lead to sampling speed acceleration, but may compromise the quality of the generated samples.

More recently, consistency models (Song et al., 2023) achieve a significant sampling speed boost, while maintaining the high quality in generated samples. Roughly speaking, consistency models merge a large number of consecutive steps in the original backward process by additionally training a consistency network via distillation or isolation. The distillation method requires a pretrained diffusion model, yet isolation lifts this requirement. In either ways, it suffices to deploy the consistency model for very few times or even a single time to generate a new sample.

Despite the empirical success, theoretical underpinnings of consistency models are limited. In particular, the following question is largely open:

What is the statistical error rate of consistency models for estimating the data distribution? How does it compare to the vanilla diffusion models?

[1] Department of Statistics and Data Science, Yale University, New Haven, US [2] Electrical and Computer Engineering, Princeton University, Princeton, US. Correspondence to: Zehao Dou <zehao.dou@yale.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

In this paper, we provide the first theoretical study towards a positive answer to the preceding question. Specifically, we consider both the distillation and isolation methods and establish statistical estimation rate of consistency models in terms of the Wasserstein distance. We summarize our contributions as follows:

- We formulate the training of consistency models as a Wasserstein distance minimization problem. This formulation is the first principled objective of consistency models, encompassing the practical consistency models' training proposed in Song et al. (2023).

- We establish statistical distribution estimation guarantees of consistency models trained under the distillation method. We demonstrate in Theorem 4.1 that the distribution estimation error is dominated by the score estimation error, showing that consistency models preserve the distribution estimation ability of vanilla diffusion models, but allow efficient sample generation.

- We extend our study to the isolation method, establishing analogous statistical estimation result. An $\widetilde{O}(n^{-1/d})$ statistical error rate is obtained in Theorem 4.2 without any pretraining on the score function.

These results are the first attempt to demystify consistency models from a statistical estimation perspective.

1.1 Related Work

Our work is related to the recent sampling theory of consistency models (Lyu et al., 2023), where they assume the score function as well as a multi-step backward process sampler have been accurately estimated. Our analysis, however, does not require such assumptions. In fact, we provide sample complexity bounds of ensuring these estimation errors being small. Apart from (Lyu et al., 2023), recent theoretical advances in diffusion models can be roughly categorized into sampling and statistical theories.

Sampling Theory of Diffusion Models This line of works show that the distribution generated by a diffusion model is close to the data distribution, as long as the score function is assumed to be accurately estimated. Specifically, De Bortoli et al. (2021); Albergo et al. (2023) study sampling from diffusion Schrödinger bridges with L_∞ accurate score functions. Concrete sampling distribution error bounds of diffusion models are provided in Block et al. (2020); Lee et al. (2022a); Chen et al. (2022); Lee et al. (2022b) under different settings, yet they all assume access to L_2 accurate score functions. Lee et al. (2022a) require the data distribution satisfying a log-Sobolev inequality. Concurrent works Chen et al. (2022) and Lee et al. (2022b) relax the log-Sobolev assumption to only having bounded moments conditions.

It is worth mentioning that Lee et al. (2022b) allow the error of the score function to be time-dependent. Recently, Chen et al. (2023c;b); Benton et al. (2023) largely enrich the study of sampling theory using diffusion models. Specifically, novel analyses based on Taylor expansions of the discretized backward process (Li et al., 2023) or localization method (Benton et al., 2023) are developed. Further, Chen et al. (2023c;b) extend to broad backward sampling methods. Besides Euclidean data, De Bortoli (2022) made the first attempt to analyze diffusion models for learning low-dimensional manifold data. Moreover, Montanari & Wu (2023) consider using diffusion processes to sample from noisy observations of symmetric spiked models and El Alaoui et al. (2023) study polynomial-time algorithms for sampling from Gibbs distributions based on diffusion processes.

Statistical Theory of Diffusion Models Distribution estimation bounds of diffusion models are first explored in Song et al. (2020b) and Liu et al. (2022) from an asymptotic statistics point of view. These results do not provide an explicit sample complexity bound. Later, Oko et al. (2023) and Chen et al. (2023a) establish sample complexity bounds of diffusion models for both Euclidean data and low-dimensional subspace data. More recently, Yuan et al. (2023) study the distribution estimation of conditional diffusion models with scalar reward guidance. Mei & Wu (2023) investigate statistical properties of diffusion models for learning high-dimensional graphical models.

Notation: For a mapping $F : \mathbb{R}^D \to \mathbb{R}^d$ and a distribution $\mathcal{D}$ supported on $\mathbb{R}^D$, $F_\sharp \mathcal{D}$ stands for the push forward distribution, which means $\mathrm{Law}(F(\mathbf{x}))$ where $\mathbf{x} \sim \mathcal{D}$. For brevity, we denote $c\mathcal{D} := f_\sharp \mathcal{D}$ where $f : \mathbf{x} \mapsto c\mathbf{x}$ is a scaling function. For two distributions $\mathcal{D}_1, \mathcal{D}_2$ supported on $\mathbb{R}^d$, denote $\mathcal{D}_1 \star \mathcal{D}_2$ as their convolution, which stands for $\mathrm{Law}(\mathbf{x}+\mathbf{y})$ where $\mathbf{x} \sim \mathcal{D}_1$ and $\mathbf{y} \sim \mathcal{D}_2$. The given dataset is $\{\mathbf{x}^j\}_{j\in[n]}$, which is assumed to be i.i.d sampled from p_{data}, our target distribution. The empirical distribution is denoted as $\widehat{p_{\mathrm{data}}} = \frac{1}{n}\sum_j \delta_{\mathbf{x}^j}$. Here, $\delta_{\mathbf{x}}$ stands for the Dirac delta distribution at point $\mathbf{x}$.

2 Diffusion Model Preliminary

We adopt a continuous time description of diffusion models, which provide rich interpretations. In practice, a proper discretization is applied accordingly. Diffusion models consist of two coupled processes. In the forward process, we gradually add noise to data following a stochastic differential equation:

$$\mathrm{d}\mathbf{x}_t = \mu(\mathbf{x}_t, t)\mathrm{d}t + \sigma(t)\mathrm{d}\mathbf{W}_t \quad \text{for } t \in [0, T]. \tag{1}$$

Here $\mathbf{W}_t(\cdot)$ is the standard Brownian motion, $\mu(\mathbf{x}_t, t)$ and $\sigma(t)$ are the drift term and the diffusion term respectively, and T is a terminal time. The forward process (1) starts from $\mathbf{x}_0 \sim p_{\mathrm{data}}$, the distribution of data. At each time $t \in [0, T]$, we denote $\mathbf{x}_t \sim p_t$ as the marginal distribution of the forward process.

As shown in Anderson (1982), the forward process enjoys a time reversal, which is termed as the backward process:

$$\mathrm{d}\mathbf{x}_t = \Big[\mu(\mathbf{x}_t, t) - \sigma(t)^2 \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)\Big] \mathrm{d}t + \sigma(t) \mathrm{d}\overline{\mathbf{W}}_t. \tag{2}$$

Here $\overline{W}_t(\cdot)$ is a standard Brownian motion with time flowing backward from T to 0 and $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ is the score function. In practice, we use a score neural network to estimate the unknown score function via denoising score matching (Song & Ermon, 2019; Vincent, 2011). It is worth mentioning that (2) is not the only backward process whose solution trajectories match the distribution of the forward process. We present the following example.

A commonly used specialization of (1) is the variance preserving SDE (VP-SDE, Dhariwal & Nichol (2021)), i.e.,

$$\mathrm{d}\mathbf{x}_t = -\frac{\beta(t)}{2} \mathbf{x}_t \mathrm{d}t + \sqrt{\beta(t)} \mathrm{d}\mathbf{W}_t. \tag{3}$$

Here $\beta(t) > 0$ is the noise schedule, which is usually chosen as a linear function over t. Under VP-SDE, we have that the transition kernel $p(\mathbf{x}_t \mid \mathbf{x}_0)$ is Gaussian satisfying

$$p(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t \mid m(t)\mathbf{x}_0, \sigma(t)^2 \boldsymbol{I}), \tag{4}$$

where

$$m(t) = \exp\left(-\frac{1}{2}\int_0^t \beta(s)\mathrm{d}s\right) \quad \text{and} \quad \sigma(t)^2 = 1 - m(t)^2.$$

At the terminal time T, the marginal distribution $p_T := p(\mathbf{x}_T)$ is approximately a standard Gaussian distribution.

The corresponding backward process to (3) is

$$\mathrm{d}\mathbf{x}_t = \left[-\frac{\beta(t)}{2}\mathbf{x}_t - \beta(t)\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)\right] \mathrm{d}t + \sqrt{\beta(t)} \mathrm{d}\overline{\mathbf{W}}_t$$

Interestingly, (3) also assumes a probability ODE flow as a backward process:

$$\mathrm{d}\mathbf{x}_t = \left[-\frac{\beta(t)}{2}\mathbf{x}_t - \frac{\beta(t)}{2}\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)\right] \mathrm{d}t. \tag{5}$$

As can be seen, the transition in (5) is deterministic.

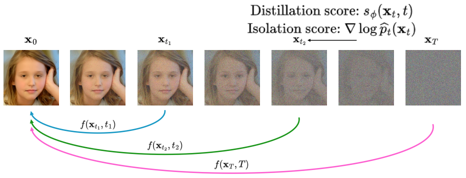
Distillation score: $s_\phi(\mathbf{x}_t, t)$
Isolation score: $\nabla \log \widehat{p}_t(\mathbf{x}_t)$
$\mathbf{x}_0$
$\mathbf{x}_{t_1}$
$\mathbf{x}_{t_2}$
$\mathbf{x}_T$
$f(\mathbf{x}_{t_1}, t_1)$
$f(\mathbf{x}_{t_2}, t_2)$
$f(\mathbf{x}_T, T)$

Figure 1. Illustration of Consistency Models: At each time step t, the consistency model $f(\cdot, t)$ will map $\mathbf{x}_t$ to $\mathbf{x}_0$ along the trajectory of probability flow ODE. We also demonstrate the score function applied at time t in both distillation training and isolation training.

3 Consistency Models Minimize Discrepancy

Consistency models merge multiple backward steps in the vanilla diffusion models to expedite the sampling. As proposed in Song et al. (2023), the training of consistency models utilizes either distillation or isolation. Unfortunately, only iterative algorithms are derived in Song et al. (2023), making the training objective elusive. In this section, we formulate the training of consistency models as a Wasserstein distance minimization problem, which encodes the original derivation in Song et al. (2023), but also enables broad modifications. To motivate the consistency models, we consider the probabilistic ODE (5) as our backward process. Consistency models seek a mapping $f_\theta(\mathbf{x}, t)$ that identifies a solution trajectory in the backward ODE to a single point. In particular, we define

$$f_\theta(\mathbf{x}, t) = \begin{cases} \mathbf{x} & t = \varepsilon \\ F_\theta(\mathbf{x}, t) & t \in (\varepsilon, T] \end{cases}, \tag{6}$$

where $F_\theta(\cdot,\cdot) : \mathbb{R}^d \times [\epsilon, T] \mapsto \mathbb{R}^d$ is a free-form deep neural network with parameter θ and ε is an early-stopping time to prevent instability (Song & Ermon, 2020). The neural network $F_\theta(\mathbf{x}, t)$ should satisfy a time-invariant property with respect to the solution trajectories in the ODE (5). Specifically, for any two time points $t_1 \neq t_2 \in [\varepsilon, T]$, we denote the contemporary generated samples as $\mathbf{x}_{t_1}$ and $\mathbf{x}_{t_2}$. Then in the ideal case, it holds that $F_\theta(\mathbf{x}_{t_1}, t_1) = F_\theta(\mathbf{x}_{t_2}, t_2) = \mathbf{x}_\varepsilon$. In other words, F_θ attempts to identify an ODE trajectory to its end point, which is the generated data point.

Training of Consistency Models The training of consistency models leverage the time-invariance of $f_\theta(\mathbf{x}, t)$. We discretize the time interval $[\varepsilon, T]$ into N uniform subintervals, with breaking points $\varepsilon = t_0 < t_1 < \ldots < t_N = T$. We denote $t_k = t_0 + k\Delta t$ where $\Delta t = \frac{T-\varepsilon}{N}$ is the length of each sub-interval. We also denote $\{\tau_k\}_{k \in [N']}$ as a subset of time steps such that $\tau_k := t_{kM}$ where $N = N'M$ with $\tau_0 = t_0 = \varepsilon$ and $\tau_{N'} = t_N = T$. Corresponding to the exposure of the time-invariance property of f_θ, consistency models aim to enforce

$$f_\theta(\cdot, \tau_k)_\sharp \mathbf{X}_{\tau_k} \overset{\text{law}}{=} f_\theta(\cdot, \tau_{k-1})_\sharp \mathbf{X}_{\tau_{k-1}} \overset{\text{law}}{=} \mathbf{X}_\varepsilon \quad \forall k \in [N'].$$

To this end, we define the following Wasserstein distance-based consistency loss for training f_θ:

$$\sum_{k=1}^{N'} W_1\left(f_\theta(\cdot, \tau_k)_\sharp \mathbf{X}_{\tau_k}, f_\theta(\cdot, \tau_{k-1})_\sharp \mathbf{X}_{\tau_{k-1}}\right). \tag{7}$$

Here $\mathbf{X}_t = \mathrm{Law}(\mathbf{x}_t) = m(t)p_{\text{data}} \star \mathcal{N}(0, \sigma(t)^2 \boldsymbol{I})$ for $\forall t \in [\varepsilon, T]$ by Equation (4). We remark that (7) accommodates to both deterministic and stochastic backward processes by measuring the distribution discrepancy, while our discussion focuses on deterministic ODEs.

Notice that the Wasserstein distance in (7) is not tractable, since we have no access to the target distribution p_{data}, let alone $\mathbf{X}_t$. Therefore, we replace it by the empirical counterpart $\widehat{p_{\text{data}}} = \frac{1}{n}\sum_j \delta_{\mathbf{x}^j}$ as well as $\mathcal{X}_t := m(t)\widehat{p_{\text{data}}} \star \mathcal{N}(0, \sigma^2(t))$, the empirical version of $\mathbf{X}_t$. We cast (7) into

$$\sum_{k=1}^{N'} W_1\left(f_\theta(\cdot, \tau_k)_\sharp \mathcal{X}_{\tau_k}, f_\theta(\cdot, \tau_{k-1})_\sharp \mathcal{X}_{\tau_{k-1}}\right). \tag{8}$$

In practice, there are two different approaches, named distillation and isolation, to determine the corresponding sample from $\mathbf{X}_{\tau_{k-1}}$ given a sample $\mathbf{x}_{\tau_k} \sim \mathbf{X}_{\tau_k}$. Both of them pushes $\mathbf{x}_{\tau_k}$ along the backward probability flow (5) by ODE update, but consistency distillation relies on a pretrained plug-in score estimator $s_\phi(\mathbf{x}, t)$ while consistency isolation does not require any pretrained models. In this work, we study both the consistency distillation and isolation in Section 4 and provide a statistical rate of $W_1\left(f_{\widehat{\theta}}(\cdot, T)_\sharp \mathcal{N}(0, \boldsymbol{I}), p_{\text{data}}\right)$ for the learned consistency model $f_{\widehat{\theta}}(\cdot, \cdot)$.

Distillation Method Given a time step τ_k as well as $\mathbf{x}_{\tau_k} \sim \mathbf{X}_{\tau_k}$, we obtain a corresponding sample $\mathbf{x}_{\tau_{k-1}}$ by running M discretization steps of probability ODE (5) solver starting from $\mathbf{x}_{\tau_k}$. For a one-step update, we denote

$$\widehat{\mathbf{x}}^{\phi}_{t_{k-1}} = \mathbf{x}_{t_k} - \Delta t \cdot \Phi(\mathbf{x}_{t_k}, t_k; \phi) := G(\mathbf{x}_{t_k}, t_k; \phi). \tag{9}$$

Here, $\Phi(\cdot, \cdot; \phi)$ is the update function of numerical ODE. In our variance preserving framework (5), we have:

$$\Phi(\mathbf{x}_{t_k}, t_k; \phi) = -\frac{\beta(t_k)}{2}\mathbf{x}_{t_k} - \frac{\beta(t_k)}{2} \cdot s_\phi(\mathbf{x}_{t_k}, t_k). \tag{10}$$

After applying M consecutive updates, we obtain $\widehat{\mathbf{x}}^{\phi,M}_{\tau_{k-1}}$ from $\mathbf{x}_{\tau_k} \sim \mathbf{X}_{\tau_k}$, which is defined as

$$\mathbf{y}_M := \mathbf{x}_{\tau_k} = \mathbf{x}_{t_{kM}},\ \mathbf{y}_{j-1} := G\left(\mathbf{y}_j, t_{(k-1)M+j}; \phi\right)$$

for $j \in [M]$, and eventually $\widehat{\mathbf{x}}^{\phi,M}_{\tau_{k-1}} := \mathbf{y}_0$. Here, the update function $G(\cdot, \cdot; \phi)$ is the same as Equation (9). For simplicity, it is equivalent to express it as

$$\begin{aligned}\widehat{\mathbf{x}}^{\phi,M}_{\tau_{k-1}} &= G_{(M)}(\mathbf{x}_{\tau_k}, \tau_k; \phi) \\ &:= G(\cdot, t_{(k-1)M+1}; \phi) \circ \ldots \circ G(\cdot, t_{kM}; \phi)(\mathbf{x}_{\tau_k}).\end{aligned}$$

In this way, we can approximate distribution $\mathbf{X}_{\tau_{k-1}}$ with $G_{(m)}(\cdot, \tau_k; \phi)_\sharp \mathbf{X}_{\tau_k}$, whose error only comes from the discretization loss of ODE solver as well as the score estimation loss of $s_\phi(\cdot, t)$. Now, we have the training objective of consistency models as follows:

$$\mathcal{L}^N_{\text{CD}}(\theta; \phi) = \sum_{k=1}^{N'} W_1\left(f_\theta(\cdot, \tau_k)_\sharp \mathcal{X}_{\tau_k}, f_\theta(\cdot, \tau_{k-1})_\sharp \widehat{\mathcal{X}}^{\phi,M}_{\tau_{k-1}}\right). \tag{11}$$

Here, $\widehat{\mathcal{X}}^{\phi,M}_{\tau_{k-1}} = G_{(M)}(\cdot, \tau_k; \phi)_\sharp \mathcal{X}_{\tau_k}$ is the underlying distribution of $\widehat{\mathbf{x}}^{\phi,M}_{\tau_{k-1}} = G_{(M)}(\mathbf{x}_{\tau_k}, \tau_k; \phi)$ where $\mathbf{x}_{\tau_k} \sim \mathcal{X}_{\tau_k}$. Our consistency model $f_{\widehat{\theta}}$ is optimized over function class $\text{Lip}(R)$, with regard to the optimization problem:

$$\widehat{\theta} = \arg \min_{\theta:\ f_\theta \in \text{Lip}(R)} \mathcal{L}^N_{\text{CD}}(\theta; \phi). \tag{12}$$

Here, $\text{Lip}(R)$ denotes the set of functions $f(\mathbf{x}, t)$ such that $f(\cdot, t)$ is Lipschitz-R continuous over $\mathbf{x}$ at any given time step $t \in [\varepsilon, T]$ with boundary condition $f(\cdot, \varepsilon) = \text{id}$.

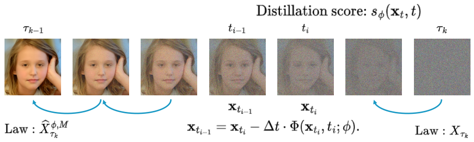
Distillation score: $s_\phi(\mathbf{x}_t, t)$
τ_{k-1}
t_{i-1}
t_i
τ_k
$\mathbf{x}_{t_{i-1}}$
$\mathbf{x}_{t_i}$
Law : $\widehat{X}^{\phi,M}_{\tau_k}$
$\mathbf{x}_{t_{i-1}} = \mathbf{x}_{t_i} - \Delta t \cdot \Phi(\mathbf{x}_{t_i}, t_i; \phi)$.
Law : X_{τ_k}

Figure 2. Illustration of $\widehat{X}^{\phi,M}_{\tau_k}$: When starting from distribution X_{τ_k} at time τ_k and following the discrete distillation-based backward process, it ends at τ_{k-1} with underlying law $\widehat{X}^{\phi,M}_{\tau_k}$.

Isolation Method Besides training in distillation, consistency models can also be trained without a pre-learned score estimator. Instead of using score model $s_\phi(\mathbf{x}_t, t)$ to approximate the true score function $\nabla \log p_t(\mathbf{x}_t)$, we can also use the following Tweedie's formula

$$\nabla \log p_t(\mathbf{x}_t) = -\mathbb{E}\left[\frac{\mathbf{x}_t - m(t)\mathbf{x}_0}{\sigma(t)^2}\middle|\mathbf{x}_t\right]$$

where $\mathbf{x}_0 \sim p_{\text{data}}$ and $p(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_0, \sigma(t)^2\boldsymbol{I})$. Since p_{data} is intractable, we make an unbiased approximation as follows:

$$\nabla \log p_t(\mathbf{x}_t) \approx -\mathbb{E}_{\mathbf{x}_0 \sim \widehat{p_{\text{data}}}}\left[\frac{\mathbf{x}_t - m(t)\mathbf{x}_0}{\sigma(t)^2}\middle|\mathbf{x}_t\right]. \tag{13}$$

Lemma 3.1. *For the approximator above, it exactly equals to the score function of distribution $\mathcal{X}_t$, i.e.*

$$-\mathbb{E}_{\mathbf{x}_0 \sim \widehat{p_{\text{data}}}}\left[\frac{\mathbf{x}_t - m(t)\mathbf{x}_0}{\sigma(t)^2}\middle|\mathbf{x}_t\right] = \nabla \log \widehat{p}_t(\mathbf{x}_t).$$

Here, $\widehat{p}_t(\cdot)$ is the density of $\mathcal{X}_t = m(t)\widehat{p_{\text{data}}} \star \mathcal{N}(0, \sigma(t)^2\boldsymbol{I})$, which is a mixture of Gaussian. Therefore, it has explicit formulation and needs no additional training.

Proof. Detailed proof is left in Appendix §A.1. □

Lemma 3.1 concludes that, taking a backward ODE step in the isolation setting is equivalent to moving along the following empirical backward diffusion ODE.

$$\text{Forward: } \mathrm{d}\mathbf{x}_t = -\frac{\beta(t)}{2}\mathbf{x}_t \mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}\mathbf{W}_t,\ \mathbf{x}_0 \sim \widehat{p_{\text{data}}}.$$

$$\text{Backward: } \mathrm{d}\mathbf{x}_t = \left[-\frac{\beta(t)}{2}\mathbf{x}_t - \frac{\beta(t)}{2}\nabla_{\mathbf{x}_t} \log \widehat{p}_t(\mathbf{x}_t)\right]\mathrm{d}t. \tag{14}$$

Its only differences with the diffusion model introduced in distillation training is that $\mathbf{x}_0 \sim \widehat{p_{\text{data}}}$ instead of p_{data}, and the empirical score function $\nabla \log \widehat{p}_t(\cdot)$ is applied instead of true score $\nabla \log p_t(\cdot)$. Under this forward SDE, it's obvious that $\text{Law}(\mathbf{x}_t) = \mathcal{X}_t$. In this case, a one-step update of the backward probability ODE at $\mathbf{x}_{t_k} \sim \mathcal{X}_{t_k}$ accurately links $\mathcal{X}_{t_k}$ to $\mathcal{X}_{t_{k-1}}$. Therefore, the isolation training objective of consistency models is as follows:

$$\mathcal{L}^N_{\text{CT}}(\theta) = \sum_{k=1}^{N'} W_1\left(f_\theta(\cdot, \tau_k)_\sharp \mathcal{X}_{\tau_k}, f_\theta(\cdot, \tau_{k-1})_\sharp \mathcal{X}_{\tau_{k-1}}\right). \tag{15}$$

Similarly, our consistency model $f_{\widehat{\theta}}$ is optimized with regard to the optimization problem:

$$\widehat{\theta} = \arg \min_{\theta:\ f_\theta \in \mathrm{Lip}(R)} \mathcal{L}_{\mathrm{CT}}^{N}(\theta). \tag{16}$$

Notice that, there is no parameter ϕ in the objective since the isolation training does not need the pre-trained score model $s_\phi(\cdot,\cdot)$.

Connection to Original Consistency Model Training in Song et al. (2023) For the practical training of consistency models, (Song et al., 2023) proposed the following sample-based consistency loss:

$$\mathcal{L}(\theta,\theta^{-};\phi) = \mathbb{E}\left[\lambda(t_k)\cdot d\left(f_\theta(\mathbf{x}_{t_k},t_k), f_{\theta^{-}}(\widehat{\mathbf{x}}^{\phi}_{t_{k-1}},t_{k-1})\right)\right].$$

Here, the expectation is taken over $k \sim \mathrm{Unif}[1,N]$ and $\mathbf{x}_t \sim \mathcal{X}_t$ for $\forall t \in [0,T]$. θ^{-} is the running average of the past values of θ in previous iterations during the optimization, and $d(\cdot,\cdot)$ is a metric function over the sample space. Besides, $\lambda(\cdot)$ is a positive weighting function over time and $\widehat{\mathbf{x}}^{\phi}_{t_{k-1}}$ is obtained by making a discretization step through backward probability flow (Equation (9)) from $\mathbf{x}_{t_k}$. In comparison, we make the following minor modifications for convenience of theoretical analysis on the statistical rate of consistency models.

We let $\lambda(\cdot) \equiv 1$, which is applied both practically and theoretically (Lyu et al., 2023). A simplification of $\theta^{-} = \theta$ is made since the optimization techniques while learning consistency models is not what we consider from the statistical point of view. We also extend the one-step ODE solver to multi-step ODE solver, which pushes $\mathbf{x}_{\tau_k}$ back to $\widehat{\mathbf{x}}^{\phi,M}_{\tau_{k-1}}$.

Another main difference is that we use Wasserstein-1 metric $W_1(\cdot,\cdot)$ over distribution space instead of the sample-based metric $d(\cdot,\cdot)$ as the training objective of consistency models. Our ultimate goal is to upper bound the distance between p_{data} and $f_{\widehat{\theta}}(\cdot,T)_\sharp\mathcal{N}(0,\boldsymbol{I})$, which makes the distribution-based metric sufficient for our analysis.

In specific, consistency models aim to learn a direct transformation f_θ that matches the distribution generated via multiple backward steps. To achieve this goal, the original training loss of consistency models requires pointwise alignment between the outputs of f_θ and multiple backward steps, such as using the l_2-distance. Nonetheless, W_1-distance is a discrepancy measure in the distributional sense and ensures the outputs of f_θ matches that of multiple backward steps in distribution. Note that W_1-distance is weaker than the pointwise l_2-distance: A small l_2-distance implies a small W_1-distance. Therefore, our analysis is derived under weaker conditions and covers the stronger l_2-distance. In practice, l_2-distance is used due to its straightforward implementation.

4 Statistical Rates of Consistency Models

In this section, we propose our main theorems for the statistical error rates of consistency models, under both settings of distillation training and isolation training.

Consistency Distillation After obtaining the global optima $f_{\widehat{\theta}}$ in the optimization problem (12), we first construct a baseline consistency model $f_{\theta^*}(\cdot,t)$ induced by natural probability flow ODE solver, named as DDPM solver, whose formulation is presented below. Next, we can upper bound the gap between these two one-step consistency models by applying the optimality condition, with the performance gap represented as

$$W_1\left(f_{\widehat{\theta}}(\cdot,T)_\sharp\mathbf{X}_T, f_{\theta^*}(\cdot,T)_\sharp\mathbf{X}_T\right).$$

Furthermore, we conclude our main theorem which upper bounds the following statistical error

$$\mathcal{L}(\widehat{\theta}) = W_1\left(f_{\widehat{\theta}}(\cdot,T)_\sharp\mathcal{N}(0,\boldsymbol{I}), p_{\mathrm{data}}\right). \tag{17}$$

by using the bounds of $W_1(\mathbf{x}_T,\mathcal{N}(0,\boldsymbol{I}))$ and the approximation error $W_1(f_{\theta^*}(\cdot,T)_\sharp\mathbf{X}_T,\mathbf{X}_\varepsilon)$ of the DDPM solver $f_{\theta^*}(\cdot,\cdot)$. Here, the baseline DDPM solver $f_{\theta^*}(\cdot,\cdot)$ is structured as an N-layer ResNet (He et al., 2016) with an inserted pretrained score estimator $s_\phi(\cdot,\cdot)$:

$$f_{\theta^*}(\mathbf{x},t) = f_{\theta^*}\left(\widehat{\mathbf{x}}^{\phi},\ t-\Delta t\right) \quad \forall t \in [t_1,T]$$

where $\widehat{\mathbf{x}}^{\phi} := \mathbf{x} + \left(\frac{\beta(t)}{2}\mathbf{x} + \frac{\beta(t)}{2}s_\phi(\mathbf{x},t)\right)\Delta t$ is a single-step numerical ODE update from $\mathbf{x}$ at time t. For $\forall t \in [\varepsilon,t_1]$, we let

$$f_{\theta^*}(\mathbf{x},t) = \mathbf{x} + \left(\frac{\beta(t)}{2}\mathbf{x} + \frac{\beta(t)}{2}s_\phi(\mathbf{x},t)\right)\cdot(t-\varepsilon).$$

This structure naturally assures that $f_{\theta^*}(\cdot,\varepsilon) = \mathrm{id}$, which makes extra reparameterization techniques unnecessary.

In this work, we propose an upper bound for the statistical rate of consistency error (17), given a pretrained score estimator s_ϕ and a global optimal solution $\widehat{\theta}$. Formally, we state our assumptions and main theorem as follows.

Assumption 4.1 (Gaussian tail)**.** For the target distribution p_{data}, it is twice continuously differentiable and it has a Gaussian tail, i.e. there exists positive constants $\alpha_1,\alpha_2 > 0$ such that

$$\mathbb{P}_{X\sim p_{\mathrm{data}}}\left[\|X\|_2 \geqslant R_0\right] \leqslant \mathbb{P}_{Z\sim\mathcal{N}(0,I)}\left[\|Z\|_2 \geqslant \frac{R_0-\alpha_1}{\alpha_2}\right]$$

holds for all $R_0 > \alpha_1$. Notice that, this assumption directly leads to the finite second order moment of p_{data}:

$$\mathcal{M}_2^2 = \mathbb{E}_{X\sim p_{\mathrm{data}}}\|X\|_2^2 < \infty.$$

As we know, the Sub-Gaussian tail is a very mild assumption, encapsulating various practical distributions, such as those with compact support set. Sub-Gaussian tail is also widely studied in existing literature on high-dimensional statistics (Wainwright, 2019).

Assumption 4.2 (Lipschitz score function)**.** For any time step $t \in [0, T]$, the score function $\nabla \log p_t(\cdot)$ is L-Lipschitz.

The two assumptions above are mild and have been widely used in relevant works (Lyu et al., 2023; Block et al., 2020; Lee et al., 2022a;b). Unlike Block et al. (2020); De Bortoli et al. (2021); Lee et al. (2022a), we do not need extra conditions on the target distribution such as log-Sobolev inequality or log-concavity, but the Gaussian tail condition is stronger than a bounded second order moment. In this paper, Assumption 4.1 is necessary since we need to bound Wasserstein distance with KL divergence. Besides, we can also remove the Lipschitz assumption on the score function by adapting analysis in Benton et al. (2023). However, it is only used for technical convenience in bounding the discretization error in Theorems 4.1 and 4.2, which is only a lower-order term.

Assumption 4.3 (Lipschitz continuity of f_{θ^*})**.** We assume that the baseline consistency model $f_{\theta^*}(\cdot, t)$ is R-Lipschitz continuous for $\forall t \in [\varepsilon, T]$.

Remark 4.1. As Caffarelli (1992) proposes, for two distributions μ and ρ with α-Hölder densities and convex support set, there exists a transformation T^* which is $(\alpha+1)$-Hölder smooth, such that $T^*_\sharp \rho = \mu$. This conclusion shows us the existence of transformation with regularity. Assumption 4.3 is natural and has been previously used in Assumption 5 of Lyu et al. (2023) and Theorem 1 of Song et al. (2023).

Remark 4.2. Notice that the existence of $f_{\theta^*}(\cdot, t)$ does not imply an access to it. Indeed, $f_{\theta^*}(\cdot, t)$ is induced by a continuous-time ODE, which is nearly impossible to be queried exactly. Therefore, we need to learn f_{θ^*} during the training of consistency models. Our Assumption 4.3 only asserts that $f_{\theta^*}(\cdot, t)$ is R-Lipschitz continuous. However, we do not have access to the ground truth θ^*.

Assumption 4.4 (Bounded coefficient)**.** In our variance preserving SDE (3), the coefficient function $\beta(t)$ is upper and lower bounded by $\overline{\beta}$ and $\underline{\beta}$, such that:

$$\underline{\beta} \leqslant \beta(t) \leqslant \overline{\beta} < \frac{1}{d \log n + d^2 \log(d/\varepsilon)} \quad \text{for } \forall t \in [\varepsilon, T].$$

Compared with Lyu et al. (2023), we do not require additional assumptions on score estimation error or consistency loss. Actually, bounding these two losses are important parts of our proof. Now, we introduce our main theorem.

Theorem 4.1 (Main Theorem 1: Distillation)**.** *Under Assumptions 4.1 - 4.4, there exists a score estimator $s_\phi(\cdot, t)$ such that the consistency model $f_{\widehat{\theta}}(\cdot, t)$ obtained from (12) satisfies that:*

$$\mathbb{E}\left[W_1\left(f_{\widehat{\theta}}(\cdot, T)_\sharp \mathcal{N}(0, \boldsymbol{I}), p_{\text{data}}\right)\right]$$
$$\lesssim \sqrt{d}R\exp(-\underline{\beta}T/2) + \frac{R\overline{\beta}dLT}{\sqrt{M}} + 6RN'n^{-1/d}$$
$$+ R\overline{\beta}\sqrt{d}\varepsilon_{\text{score}} \cdot \sqrt{\frac{TN'}{\varepsilon}} + \sqrt{d\overline{\beta}\varepsilon},$$

where R is the Lipschitz constraint of the optimization problem (12), and the expectation is taken with respect to the choice of dataset $\{\mathbf{x}^j\}_{j\in[n]}$. $\varepsilon_{\text{score}} = \mathcal{O}(n^{-1/(d+5)})$ stands for the score estimation error.

We interpret error terms in Theorem 4.1 as follows. $\sqrt{d}R\exp(-\underline{\beta}T/2)$ represents the convergence error of the forward process. $\frac{R\overline{\beta}dLT}{\sqrt{M}}$ is the discretization error of ODE updates. $6RN'n^{-1/d}$ represents the concentration gap. $R\overline{\beta}\sqrt{d}\varepsilon_{\text{score}} \cdot \sqrt{\frac{TN'}{\varepsilon}}$ is the score estimation error and $\sqrt{d\overline{\beta}\varepsilon}$ is the error caused by early stopping. We show in the following remark that the dominating error term is the score estimation error, with proper choice of hyperparameters.

Remark 4.3. After picking $\overline{\beta}, \underline{\beta} \asymp \frac{1}{d \log n}$, $T = (\log n)^3$, $M = d^2 n^{\frac{1}{d+5}}, N' = \log n$ and $\varepsilon = \sqrt{TN'}n^{-\frac{1}{d+5}} = \log^2 n \cdot n^{-\frac{1}{d+5}}$, we have the bound:

$$\mathbb{E}\left[W_1\left(f_{\widehat{\theta}}(\cdot, T)_\sharp \mathcal{N}(0, \boldsymbol{I}), p_{\text{data}}\right)\right] \lesssim \sqrt{\log n} \cdot n^{-\frac{1}{2(d+5)}}.$$

Now, we obtain a $\widetilde{\mathcal{O}}\left(n^{-\frac{1}{2(d+5)}}\right)$ bound for the Wasserstein estimation error of consistency model via distillation, preserving the distribution estimation rate of the vanilla diffusion models as shown in Chen et al. (2023a). This indicates that consistency models maintain the quality of the generated samples, while allowing fast sampling.

Remark 4.4. We adopt the nonparametric statistics point of view and the score estimation rate highlights an exponential dependence on the dimension d, which is in fact optimal without further assumptions. Nonetheless, it can be reduced in multiple ways: (1) Practical data has rich low-dimensional structures, which is a critical reason why practical diffusion models can be effectively trained. As shown in Chen et al. (2023a), when data has intrinsic subspace structures, the score estimation error only depends on the intrinsic dimension. (2) In parametric settings, we can even obtain a score estimation rate in the order of $\text{poly}(d)/\sqrt{n}$ (Yuan et al., 2024). We remark that in both cases, the improved convergence rate is tied to data structure assumptions.

Consistency Isolation Similar to the consistency distillation case, we still need to construct a baseline consistency model $f_{\theta^*}(\cdot, \cdot)$, named as empirical DDPM solver, which replaces the inserted pretrained score model $s_\phi(\mathbf{x}, t)$ with $\nabla \log \widehat{p}_t(\mathbf{x})$, the explicit score of a mixture of Gaussian:

$$f_{\theta^*}(\mathbf{x}, t) = f_{\theta^*}(\widehat{\mathbf{x}}, t - \Delta t) \quad \forall t \in [t_1, T]$$

where $\widehat{x} := \mathbf{x} + \left(\frac{\beta(t)}{2}\mathbf{x} + \frac{\beta(t)}{2}\nabla \log \widehat{p}_t(\mathbf{x})\right)\Delta t$ is a single-step numerical ODE update from $\mathbf{x}$ at time t along the empirical backward ODE (14). For $\forall t \in [\varepsilon, t_1]$, we set

$$f_{\theta^*}(\mathbf{x}, t) = \mathbf{x} + \left(\frac{\beta(t)}{2}\mathbf{x} + \frac{\beta(t)}{2}\nabla \log \widehat{p}_t(\mathbf{x})\right) \cdot (t - \varepsilon).$$

After obtaining $f_{\widehat{\theta}}$ from the optimization problem (16), we upper bound the performance gap induced by learned one-step consistency model $f_{\widehat{\theta}}(\cdot, T)$ and the empirical DDPM solver $f_{\theta^*}(\cdot, T)$, which is evaluated by $W_1\left(f_{\widehat{\theta}}(\cdot, T)_\sharp \mathcal{X}_T, f_{\theta^*}(\cdot, T)_\sharp \mathcal{X}_T\right)$. Furthermore, it leads to our main theorem on consistency isolation, which upper bounds the statistical error

$$\mathcal{L}(\widehat{\theta}) = W_1\left(f_{\widehat{\theta}}(\cdot, T)_\sharp \mathcal{N}(0, \boldsymbol{I}), p_{\text{data}}\right).$$

To achieve this result, we require a stronger version of Assumption 4.1, that the target distribution p_{data} has a bounded support set:

Assumption 4.5 (Bounded support set)**.** The target distribution p_{data} has a bounded support set such that:

$$\mathbb{P}_{X \sim p_{\text{data}}}\left[\|X\|_2 \leqslant R_0\right] = 1.$$

Here, we require a much stronger assumption than the Gaussian tail because a Lipschitz continuity condition is needed over the empirical score function $\nabla \log \widehat{p}_t(\cdot)$ for $\forall t \in [\varepsilon, T]$ to replace Assumption 4.2. Now, we state our main theorem on consistency isolation as follows:

Theorem 4.2 (Main Theorem 2: Isolation)**.** *Under Assumptions 4.3- 4.5, the consistency model $f_{\widehat{\theta}}(\cdot, t)$ obtained from Equation (16) satisfies that:*

$$\begin{aligned}&\mathbb{E}\left[W_1\left(f_{\widehat{\theta}}(\cdot, T)_\sharp \mathcal{N}(0, \boldsymbol{I}), p_{\text{data}}\right)\right]\\ &\lesssim \sqrt{d} R \exp\left(-\underline{\beta} T/2\right) + R n^{-1/d} + \frac{d \overline{\beta} R_0^2 T}{\underline{\beta}^2 \varepsilon^2 \sqrt{M}} + \sqrt{d \overline{\beta} \varepsilon},\end{aligned}$$

where R is the Lipschitz constraint of the optimization problem (16), and the expectation is taken over the dataset.

Remark 4.5. After picking $\overline{\beta}, \underline{\beta} \asymp \frac{1}{d \log n}$, $\varepsilon = n^{-2/d}$, $T = d(\log n)^3$, $M = d^2(\log n)^8 \cdot n^{10/d}$, we have the bound:

$$\mathbb{E}\left[W_1\left(f_{\widehat{\theta}}(\cdot, T)_\sharp \mathcal{N}(0, \boldsymbol{I}), p_{\text{data}}\right)\right] \lesssim n^{-1/d}.$$

Now, we obtain a $\widetilde{\mathcal{O}}\left(n^{-1/d}\right)$ bound for the Wasserstein estimation error of consistency model via isolation. Note that the rate of convergence is not directly comparable to distillation method, due to the distinct training procedure.

Remark 4.6. Assumption 4.5 can be straightforwardly relaxed to the sub-Gaussian tail assumption (Assumption 4.1) since the tail shrinks exponentially fast under the sub-Gaussian assumption, which makes it plausible to truncate the data domain with well-controlled truncation errors, and then our analysis reduces to the bounded support case.

5 Proof Sketch for Consistency Distillation

In this section, we provide the proof sketch for the main theorems proposed in the previous part. First, we propose an overview of the entire proof sketch.

5.1 Technical Overview

We now present a detailed technical overview for the proof of the statistical error rate for distillation consistency models (Theorem 4.1). For the proof of isolation consistency models (Theorem 4.2), it follows very similar ideas and we leave the detailed proof in Appendix §D.

As we state above, our ultimate goal is to upper bound the statistical estimation error $W_1(f_{\widehat{\theta}}(\cdot, T)_\sharp \mathcal{N}(0, \boldsymbol{I}), p_{\text{data}})$, the distance between the true distribution p_{data} and standard Gaussian pushed forward by our learned one-step consistency model $f_{\widehat{\theta}}(\cdot, T)$ by distillation. To achieve this, we construct a DDPM solver $f_{\theta^*}(\cdot, \cdot)$ which is assumed to be R-Lipschitz continuous at all time steps $t \in [\varepsilon, T]$, and upper bound the performance gap between $f_{\widehat{\theta}}$ and f_{θ^*}.

In the first step, we study the approximation properties of score estimation, and our purpose is to show the existence of a score network $s_\phi(\mathbf{x}, t)$ with small approximation error $\mathbb{E}\|s_\phi(\mathbf{x}, t) - \nabla \log p_t(\mathbf{x})\|^2$. With the score approximation error bounded, we can conclude the proximity between the true backward probability ODE and that with pretrained score model inserted.

Next, we aim to bound the performance gap between the learned one-step consistency model $f_{\widehat{\theta}}$ and the DDPM solver f_{θ^*}. According to the training objective (12) as well as Assumption 4.3 which makes f_{θ^*} also included in the constraint set $\text{Lip}(R)$, we can apply the optimality inequality

$$\mathcal{L}_{\text{CD}}^N(\widehat{\theta}; \phi) \leqslant \mathcal{L}_{\text{CD}}^N(\theta^*; \phi) \tag{18}$$

Through some mathematical calculation, we show that the performance gap is directly relevant to the concentration gap between empirical and population distributions, as well as the error caused by the numerical ODE update. There are two main types of error taking place during the ODE update, which are the discretization error and the score estimation error. The former one is directly relevant to the length of time sub-intervals Δt while the bound of the latter one is already solved in our first step.

After that, we finally come to our main theorem upper bounding the statistical error (17). It can be smoothly obtained by combining the performance gap, the tail bounds $W_1(\mathbf{X}_T, \mathcal{N}(0, \boldsymbol{I}))$, $W_1(\mathbf{X}_\varepsilon, p_{\text{data}})$ as well as the estimation error of the DDPM solver $W_1(f_{\theta^*}(\cdot, T)_\sharp \mathbf{X}_T, \mathbf{X}_\varepsilon)$.

In contrast, for the proof of Theorem 4.2, the major difference in the isolation setting is that there is no score estimation error since the isolation training does not involve any pretrained score models. For each ODE update, discretization is the only error that takes place. Another technical difficulty is to guarantee the Lipschitz continuity of the empirical score functions involved in the backward process.

5.2 Approximation Error for Score Estimation

In Chen et al. (2023a), the authors introduce the l-layer ReLU neural network class $\text{NN}(l, M, J, K, \kappa, \gamma, \gamma_t)$ as follows and propose a score approximation error, with the result shown in Lemma C.1. After removing the properties we do not need, we can make the following conclusion:

Lemma 5.1. *There exists a score estimator function $s_\phi(\cdot,\cdot)$ in the class of neural networks, such that: (1) $s_\phi(\cdot,t)$ is L_{score}-Lipschitz continuous for any given $t\in[\varepsilon,T]$ where $L_{\text{score}}=\mathcal{O}(10d(1+L))$; (2) $\|s_\phi(\mathbf{x},t)\|_2\leqslant U_{\text{score}}$ holds for $\forall\mathbf{x}\in\mathbb{R}^d,t\in[\varepsilon,T]$ where $U_{\text{score}}=\mathcal{O}(2d\log n+2d^2\log(d/\varepsilon))$; (3) The mean integrated squared error can be upper bounded by:*

$$\frac{1}{t_b-t_a}\int_{t_a}^{t_b}\|s_\phi(\cdot,t)-\nabla\log p_t(\cdot)\|^2_{L^2(\mathbf{X}_t)}\,\mathrm{d}t = \widetilde{O}\left(\frac{1}{\varepsilon}n^{-\frac{2}{d+5}}\right)\quad\forall\varepsilon\leqslant t_a<t_b\leqslant T.$$

According to this result, we can use the method of induction to provide a loose upper bound for the Lipschitz constant of $f_{\theta^*}(\cdot,\cdot)$, the baseline DDPM solver with score estimator $s_\phi(\cdot,\cdot)$ injected.

Corollary 5.1. *Assume the information decay rate $\beta(t)$ in Equation (3) is bounded as $\underline{\beta}\leqslant\beta(t)\leqslant\overline{\beta}$ for $\forall t\in[\varepsilon,T]$, then the trivial upper bound for the Lipschitz constant of $f_{\theta^*}(\cdot,t)$ is $\exp(Cd\overline{\beta}T)$ for any given t. Here $C=10(1+L)$ is a pure constant.*

Proof. Detailed proof is left in Appendix §C.2. □

By Lemma 5.1, we get the approximation error bound of score model, which is part of the performance gap between $f_{\widehat{\theta}}$ and f_{θ^*}. In the next part, we apply the optimality inequality (18) and decompose the consistency loss into several error terms which are easier to analyze. We will also show that these terms stand for the concentration gap and the numerical ODE update error.

5.3 Upper Bound the Consistency Loss

According to the structure of f_{θ^*}, we have

$$f_{\theta^*}(\cdot,\tau_k)=f_{\theta^*}(\cdot,\tau_{k-1})\circ G_{(M)}(\cdot,\tau_k;\phi).$$

Denote $\widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}:=G_{(M)}(\cdot,\tau_k;\phi)$ as the underlying distribution of $\widehat{\mathbf{x}}^{\phi,M}_{\tau_{k-1}}=G_{(M)}(\mathbf{x}_{\tau_k},\tau_k;\phi)$ where $\mathbf{x}_{\tau_k}\sim\mathbf{X}_{\tau_k}$, then it holds by definition that:

$$f_{\theta^*}(\cdot,\tau_k)_\sharp\mathbf{X}_{\tau_k}\overset{\text{law}}{=}f_{\theta^*}(\cdot,\tau_{k-1})_\sharp\widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}\quad\forall k\in[N'].\tag{19}$$

This equation lays the foundation of recursive analysis between adjacent time steps. After Combining with the optimality inequality (18), we can decompose the performance gap between $f_{\widehat{\theta}}$ and f_{θ^*} (also known as consistency loss) into four loss terms, which is shown in the following lemma.

Lemma 5.2. *We can upper bound the consistency loss as:*

$$W_1\left(f_{\widehat{\theta}}(\cdot,T)_\sharp\mathbf{X}_T,f_{\theta^*}(\cdot,T)_\sharp\mathbf{X}_T\right)\leqslant I_1+I_2+I_3+I_4.\tag{20}$$

Here, the four loss terms I_i $(1\leqslant i\leqslant4)$ have their formulations as follows:

$$I_1:=\sum_{k=1}^{N'}W_1\left(f_{\widehat{\theta}}(\cdot,\tau_{k-1})_\sharp\mathbf{X}_{\tau_{k-1}},f_{\widehat{\theta}}(\cdot,\tau_{k-1})_\sharp\widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}\right),$$

$$I_2:=\sum_{k=1}^{N'}W_1\left(f_{\theta^*}(\cdot,\tau_{k-1})_\sharp\mathbf{X}_{\tau_{k-1}},f_{\theta^*}(\cdot,\tau_{k-1})_\sharp\widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}\right),$$

$$I_3:=\sum_{k=1}^{N'}\Big[W_1\left(f_{\widehat{\theta}}(\cdot,\tau_k)_\sharp\mathbf{X}_{\tau_k},f_{\widehat{\theta}}(\cdot,\tau_{k-1})_\sharp\widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}\right) - W_1\left(f_{\widehat{\theta}}(\cdot,\tau_k)_\sharp\mathcal{X}_{\tau_k},f_{\widehat{\theta}}(\cdot,\tau_{k-1})_\sharp\widehat{\mathcal{X}}^{\phi,M}_{\tau_{k-1}}\right)\Big]$$

$$I_4:=\sum_{k=1}^{N'}\Big[W_1\left(f_{\theta^*}(\cdot,\tau_k)_\sharp\mathcal{X}_{\tau_k},f_{\theta^*}(\cdot,\tau_{k-1})_\sharp\widehat{\mathcal{X}}^{\phi,M}_{\tau_{k-1}}\right) - W_1\left(f_{\theta^*}(\cdot,\tau_k)_\sharp\mathbf{X}_{\tau_k},f_{\theta^*}(\cdot,\tau_{k-1})_\sharp\widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}\right)\Big].$$

Proof. Detailed proof is left in Appendix §C.3. □

As we can see, both I_1,I_2 show the multi-step discretization error of ODE solver and both I_3,I_4 show the concentration gap between empirical and population Wasserstein-1 distances. Next, we start with I_1,I_2. The technical difficulties on upper bound these terms come from two aspects. One is to bound the KL divergence between the true ODE flow measure and that with pretrained score function $s_\phi(\cdot,\cdot)$ inserted. The other is to bound Wasserstein distance with KL divergence, which is impossible in general but achievable under Gaussian tail condition (Assumption 4.1). After overcoming these obstacles, we prove the following lemma.

Lemma 5.3. *Under the Assumption 4.1-4.4, we can upper bound I_1,I_2 introduced in Lemma 5.2 as:*

$$I_1+I_2\lesssim R\overline{\beta}dL\cdot\frac{T}{\sqrt{M}}+R\overline{\beta}\sqrt{d}n^{-\frac{1}{d+5}}\cdot\sqrt{\frac{TN'}{\varepsilon}}.$$

Here, R is the Lipschitz constraint in (12).

Proof. Detailed proof is left in Appendix §C.4. □

Next, we upper bound I_3,I_4. After transforming them into the Wasserstein distance between empirical and population distributions, we prove the following lemma.

Lemma 5.4. *Under the Assumption 4.1-4.4, we can upper bound I_3,I_4 introduced in Lemma 5.2 as:*

$$\mathbb{E}\left[I_3+I_4\right]\leqslant6RN'\cdot n^{-1/d}.$$

Here, the expectation is taken with respect to the randomness of dataset $\{\mathbf{x}^1,\mathbf{x}^2,\ldots,\mathbf{x}^n\}$.

Proof. Detailed proof is left in Appendix §C.5. □

Now we can combine all the results above and get:

$$\mathbb{E}\left[W_1\left(f_{\widehat{\theta}}(\cdot,T)_\sharp\mathbf{X}_T,f_{\theta^*}(\cdot,T)_\sharp\mathbf{X}_T\right)\right]\lesssim\frac{R\overline{\beta}dLT}{\sqrt{M}}+R\overline{\beta}n^{-\frac{1}{d+5}}\sqrt{\frac{dTN'}{\varepsilon}}+6RN'n^{-1/d}\tag{21}$$

holds under Assumption 4.1-4.4.

5.4 Proof of Main Theorem 1

In order to bound the statistical error $\mathcal{L}(\widehat{\theta})$ defined in (17), we still need to bound two additional loss terms: $W_1\left(f_{\widehat{\theta}}(\cdot,T)_\sharp\mathcal{N}(0,\boldsymbol{I}), f_{\theta^*}(\cdot,T)_\sharp\mathbf{X}_T\right)$ and the estimation error of DDPM solver $W_1\left(f_{\theta^*}(\cdot,T)_\sharp\mathbf{X}_T, p_{\mathrm{data}}\right)$. Since $f_{\widehat{\theta}}(\cdot,T)$ is R-Lipschitz continuous, we have

$$W_1\left(f_{\widehat{\theta}}(\cdot,T)_\sharp\mathbf{X}_T, f_{\widehat{\theta}}(\cdot,T)_\sharp\mathcal{N}(0,\boldsymbol{I})\right) \leqslant R\cdot W_1(\mathbf{X}_T,\mathcal{N}(0,\boldsymbol{I})).$$

Therefore, we first need to bound $W_1(\mathbf{X}_T,\mathcal{N}(0,\boldsymbol{I}))$ in the following lemma.

Lemma 5.5. *For the distribution $\mathbf{X}_T$, its Wasserstein distance from the standard Gaussian $\mathcal{N}(0,\boldsymbol{I})$ can be upper bounded as:*

$$W_1(\mathbf{X}_T,\mathcal{N}(0,\boldsymbol{I})) \lesssim \sqrt{d}\exp(-\underline{\beta}T/2).$$

Proof. Detailed proof is left in Appendix §C.6. □

Next, we bound $W_1\left(f_{\theta^*}(\cdot,T)_\sharp\mathbf{X}_T, p_{\mathrm{data}}\right)$, which requires an extension on the existing result on DDPM estimation error (Theorem 2 of (Chen et al., 2022)) as well as the technique of bounding Wasserstein distance with KL divergence.

Lemma 5.6. *Under Assumption 4.1-4.4, we bound the estimation error of DDPM solver as:*

$$\begin{aligned} &W_1\left(f_{\theta^*}(\cdot,T)_\sharp\mathbf{X}_T, p_{\mathrm{data}}\right) \\ &\lesssim \overline{\beta}Ld\sqrt{T\Delta t} + \overline{\beta}\sqrt{\frac{dT}{\varepsilon}}n^{-\frac{1}{d+5}} + \sqrt{d\overline{\beta}\varepsilon}. \end{aligned}$$

Proof. Detailed proof is left in Appendix §C.7. □

Now, after summing up Lemma 5.5, 5.6 and Equation (21) together, we finally come to our main theorem 4.1:

$$\begin{aligned} &\mathbb{E}\left[W_1\left(f_{\widehat{\theta}}(\cdot,T)_\sharp\mathcal{N}(0,\boldsymbol{I}), p_{\mathrm{data}}\right)\right] \lesssim \sqrt{d}R\exp(-\underline{\beta}T/2) \\ &+ \frac{R\overline{\beta}dLT}{\sqrt{M}} + R\overline{\beta}n^{-\frac{1}{d+5}}\sqrt{\frac{dTN'}{\varepsilon}} + \sqrt{d\overline{\beta}\varepsilon} + 6RN'n^{-1/d}. \end{aligned}$$

6 Conclusion

In this paper, we have provided the first statistical theory of consistency diffusion models. In particular, we have formulated the consistency models' training as a Wasserstein discrepancy minimization problem. Further, we have established sample complexity bounds for consistency models in estimating nonparametric data distributions. The obtained convergence rate closely matches the vanilla diffusion models, indicating consistency models boost the sampling speed without significantly scarifying the sample generation quality. Our analyses have covered both the distillation and isolation methods for training consistency models.

Acknowledgements

Mengdi Wang acknowledges the support by NSF IIS-2107304, NSF CPS-2312093, ONR 1006977 and Genmab.

Impact Statement

This work presents novel statistical analysis of consistency diffusion models. Our treatment reveals a hidden connection between consistency models to discrepancy measure minimization, a missing piece in understanding consistency models. Built upon our formulated analytical framework in Section 3, we anticipate broader explorations into acceleration methods for diffusion models. Meanwhile, our research opens new possibility of studying latent consistency models (Luo et al., 2023). We do not foresee any ethical concerns.

References

Albergo, M. S., Boffi, N. M., and Vanden-Eijnden, E. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Anderson, B. D. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Benton, J., De Bortoli, V., Doucet, A., and Deligiannidis, G. Linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*, 2023.

Block, A., Mroueh, Y., and Rakhlin, A. Generative modeling with denoising auto-encoders and langevin sampling. *arXiv preprint arXiv:2002.00107*, 2020.

Caffarelli, L. A. The regularity of mappings with a convex potential. *Journal of the American Mathematical Society*, 5(1):99–104, 1992.

Chen, M., Huang, K., Zhao, T., and Wang, M. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. *arXiv preprint arXiv:2302.07194*, 2023a.

Chen, N., Zhang, Y., Zen, H., Weiss, R. J., Norouzi, M., and Chan, W. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. R. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.

Chen, S., Chewi, S., Lee, H., Li, Y., Lu, J., and Salim, A. The probability flow ode is provably fast. *arXiv preprint arXiv:2305.11798*, 2023b.

Chen, S., Daras, G., and Dimakis, A. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for ddim-type samplers. In *International Conference on Machine Learning*, pp. 4462–4484. PMLR, 2023c.

Chi, C., Feng, S., Du, Y., Xu, Z., Cousineau, E., Burchfiel, B., and Song, S. Diffusion Policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023.

Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., Yosinski, J., and Liu, R. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*, 2019.

De Bortoli, V. Convergence of denoising diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2208.05314*, 2022.

De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

El Alaoui, A., Montanari, A., and Sellke, M. Sampling from mean-field gibbs measures via diffusion processes. *arXiv preprint arXiv:2310.08912*, 2023.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Gruver, N., Stanton, S., Frey, N. C., Rudner, T. G., Hotzel, I., Lafrance-Vanasse, J., Rajpal, A., Cho, K., and Wilson, A. G. Protein design with guided discrete diffusion. *arXiv preprint arXiv:2305.20009*, 2023.

Hansen-Estruch, P., Kostrikov, I., Janner, M., Kuba, J. G., and Levine, S. IDQL: Implicit Q-learning as an actor-critic method with diffusion policies. *arXiv preprint arXiv:2304.10573*, 2023.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35: 26565–26577, 2022.

Kingma, D. P., Welling, M., et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

Lee, H., Lu, J., and Tan, Y. Convergence for score-based generative modeling with polynomial complexity. *arXiv preprint arXiv:2206.06227*, 2022a.

Lee, H., Lu, J., and Tan, Y. Convergence of score-based generative modeling for general data distributions. *arXiv preprint arXiv:2209.12381*, 2022b.

Lee, J. S., Kim, J., and Kim, P. M. Proteinsgm: Score-based generative modeling for de novo protein design. *bioRxiv*, pp. 2022–07, 2022c.

Li, X., Thickstun, J., Gulrajani, I., Liang, P. S., and Hashimoto, T. B. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022.

Li, X., Ren, Y., Jin, X., Lan, C., Wang, X., Zeng, W., Wang, X., and Chen, Z. Diffusion models for image restoration and enhancement–a comprehensive survey. *arXiv preprint arXiv:2308.09388*, 2023.

Liu, X., Wu, L., Ye, M., and Liu, Q. Let us build bridges: Understanding and extending diffusion generative models. *arXiv preprint arXiv:2208.14699*, 2022.

Lovelace, J., Kishore, V., Wan, C., Shekhtman, E., and Weinberger, K. Latent diffusion for language generation. *arXiv preprint arXiv:2212.09462*, 2022.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

Luo, S., Su, Y., Peng, X., Wang, S., Peng, J., and Ma, J. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. *Advances in Neural Information Processing Systems*, 35: 9754–9767, 2022.

Luo, S., Tan, Y., Patil, S., Gu, D., von Platen, P., Passos, A., Huang, L., Li, J., and Zhao, H. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023.

Lyu, J., Chen, Z., and Feng, S. Convergence guarantee for consistency models. *arXiv preprint arXiv:2308.11449*, 2023.

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

Montanari, A. and Wu, Y. Posterior sampling from the spiked models via diffusion processes. *arXiv preprint arXiv:2304.11449*, 2023.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *Proceedings of the International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Oko, K., Akiyama, S., and Suzuki, T. Diffusion models are minimax optimal distribution estimators. *arXiv preprint arXiv:2303.01861*, 2023.

Pearce, T., Rashid, T., Kanervisto, A., Bignell, D., Sun, M., Georgescu, R., Macua, S. V., Tan, S. Z., Momennejad, I., Hofmann, K., and Devlin, S. Imitating human behaviour with diffusion models. *arXiv preprint arXiv:2301.10677*, 2023.

Reuss, M., Li, M., Jia, X., and Lioutikov, R. Goal-conditioned imitation learning using score-based diffusion policies. *arXiv preprint arXiv:2304.02532*, 2023.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Song, Y. and Ermon, S. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

Song, Y., Garg, S., Shi, J., and Ermon, S. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, pp. 574–584. PMLR, 2020b.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020c.

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Weed, J. and Bach, F. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. 2019.

Yu, P., Xie, S., Ma, X., Jia, B., Pang, B., Gao, R., Zhu, Y., Zhu, S.-C., and Wu, Y. N. Latent diffusion energy-based model for interpretable text modeling. *arXiv preprint arXiv:2206.05895*, 2022.

Yuan, H., Huang, K., Ni, C., Chen, M., and Wang, M. Reward-directed conditional diffusion: Provable distribution estimation and reward improvement. *arXiv preprint arXiv:2307.07055*, 2023.

Yuan, H., Huang, K., Ni, C., Chen, M., and Wang, M. Reward-directed conditional diffusion: Provable distribution estimation and reward improvement. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhang, Q., Tao, M., and Chen, Y. gddim: Generalized denoising diffusion implicit models. *arXiv preprint arXiv:2206.05564*, 2022.

A Proofs in Section 3

A.1 Proof of Lemma 3.1

When $\mathbf{x}_0$ follows the empirical distribution $\widehat{p_{\text{data}}} = \frac{1}{n}\sum_{j=1}^n \delta_{\mathbf{x}^j}$, then the posterior distribution $p(\mathbf{x}_0 \mid \mathbf{x}_t)$ for a given $\mathbf{x}_t \sim \mathcal{N}(m(t)\mathbf{x}_0, \sigma(t)^2\boldsymbol{I})$ can be simply represented as:

$$p(\mathbf{x}_0 = \mathbf{x}^j \mid \mathbf{x}_t) \propto \exp\left(-\frac{\|m(t)\mathbf{x}^j - \mathbf{x}_t\|^2}{2\sigma(t)^2}\right),$$

which leads to the following posterior mean:

$$\mathbb{E}_{\mathbf{x}_0\sim\widehat{p_{\text{data}}}}[\mathbf{x}_0 \mid \mathbf{x}_t] = \sum_{j=1}^n \mathbf{x}^j \cdot p(\mathbf{x}_0 = \mathbf{x}^j \mid \mathbf{x}_t) = \frac{\sum_{j=1}^n \mathbf{x}^j \cdot \exp\left(-\frac{\|m(t)\mathbf{x}^j - \mathbf{x}_t\|^2}{2\sigma(t)^2}\right)}{\sum_{j=1}^n \exp\left(-\frac{\|m(t)\mathbf{x}^j - \mathbf{x}_t\|^2}{2\sigma(t)^2}\right)}.$$

Therefore, the score function has the following unbiased estimation:

$$\begin{aligned}\nabla \log p_t(\mathbf{x}_t) &\approx -\mathbb{E}_{\mathbf{x}_0\sim\widehat{p_{\text{data}}}}\left[\frac{\mathbf{x}_t - m(t)\mathbf{x}_0}{\sigma(t)^2}\middle|\mathbf{x}_t\right] = \frac{\sum_{j=1}^n -\frac{\mathbf{x}_t - m(t)\mathbf{x}^j}{\sigma(t)^2} \cdot \exp\left(-\frac{\|m(t)\mathbf{x}^j - \mathbf{x}_t\|^2}{2\sigma(t)^2}\right)}{\sum_{j=1}^n \exp\left(-\frac{\|m(t)\mathbf{x}^j - \mathbf{x}_t\|^2}{2\sigma(t)^2}\right)} \\ &= \frac{\nabla_{\mathbf{x}_t}\sum_{j=1}^n \exp\left(-\frac{\|m(t)\mathbf{x}^j - \mathbf{x}_t\|^2}{2\sigma(t)^2}\right)}{\sum_{j=1}^n \exp\left(-\frac{\|m(t)\mathbf{x}^j - \mathbf{x}_t\|^2}{2\sigma(t)^2}\right)} = \nabla_{\mathbf{x}_t} \log\left[\frac{1}{n}\sum_{j=1}^n \exp\left(-\frac{\|m(t)\mathbf{x}^j - \mathbf{x}_t\|^2}{2\sigma(t)^2}\right)\right].\end{aligned} \tag{22}$$

Notice that, $\frac{1}{n}\sum_{j=1}^n \exp\left(-\frac{\|m(t)\mathbf{x}^j - \mathbf{x}_t\|^2}{2\sigma(t)^2}\right)$ is exactly the density of $\mathcal{X}_t = m(t)\widehat{p_{\text{data}}} \star \mathcal{N}(0, \sigma(t)^2)$.

B Some Useful Lemmas

In this section, we introduce some lemmas directly related to the Girsanov's theorem and techniques from (Chen et al., 2022). We also propose some propositions on Gaussian tails and provide a technique to upper bound Wasserstein distance with KL divergence for distributions with Gaussian tail.

Lemma B.1. *For any $k = 1, 2, \ldots, N'$, it holds that:*

$$\mathrm{KL}\left(\mathbf{X}_{\tau_k}, \widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}\right) \leqslant \sum_{i=M(k-1)+1}^{Mk} \mathbb{E}\int_{t_{i-1}}^{t_i} \frac{\beta(t)^2}{2}\left\|s_\phi(\mathbf{X}_{t_i}, t_i) - \nabla \log p_t(\mathbf{X}_t)\right\|^2 \mathrm{d}t. \tag{23}$$

Here, the expectation is taken over the forward diffusion process. Without approximating $\mathbf{X}_T$ with standard Gaussian distribution $\mathcal{N}(0, \boldsymbol{I})$, the forward diffusion and the back diffusion share the trajectory with exactly the same marginal distributions.

Next, we need to upper bound the right hand side of Equation (23). Actually, we can directly use Theorem 9 in (Chen et al., 2022) and conclude that:

Lemma B.2. *For each $k = 1, 2, \ldots, N$ and $t \in [t_{k-1}, t_k]$, it holds that:*

$$\mathbb{E}\|s_\phi(\mathbf{X}_{t_k}, t_k) - \nabla \log p_t(\mathbf{X}_t)\|^2 \lesssim \varepsilon^2_{t_k} + L^2 d\Delta t + L^2\mathcal{M}_2^2\Delta t^2$$

where $\varepsilon^2_{t_k}$ is the score estimation error at time step t_k:

$$\varepsilon^2_{t_k} = \mathbb{E}_{\mathbf{x}\sim\mathbf{X}_{t_k}}\|s_\phi(\mathbf{x}, t_k) - \nabla \log p_{t_k}(\mathbf{x})\|^2,$$

and the expectation is taken over the forward diffusion process.

After combining Lemma B.1, Lemma B.2 and the score estimation error (Lemma 5.1), it holds that: for all $k = 1, 2, \ldots, N$

$$\begin{aligned}\mathrm{KL}\left(\mathbf{X}_{\tau_k}, \widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}\right) &\leqslant \frac{\overline{\beta}^2}{2}\left(L^2 d\Delta t + L^2\mathcal{M}_2^2\Delta t^2\right) \cdot M\Delta t + \frac{\overline{\beta}^2}{2}\mathbb{E}\int_{\tau_{k-1}}^{\tau_k} \left\|s_\phi(\mathbf{x}_t, t) - \nabla \log p_t(\mathbf{x}_t)\right\|^2 \mathrm{d}t \\ &\lesssim \overline{\beta}^2\left(L^2 d\Delta t \cdot M\Delta t + \frac{1}{\varepsilon}n^{-\frac{2}{d+5}} \cdot M\Delta t\right)\end{aligned} \tag{24}$$

Another major technical result we need is to upper bound Wasserstein distance with KL divergence, which is impossible in the general case. However under Assumption 4.1, we will show that all the variables like $\mathbf{X}_t$ and $\widehat{\mathbf{X}}^{\phi,M}_t$ have Gaussian tail, which enables the upper bounding. To achieve this, we propose a rigorous notion of Gaussian tail before proving a more general result.

Lemma B.3. *For constants $c_1, c_2 > 0$, we call a d-dimensional random variable X having a (c_1, c_2)-Gaussian tail if there exists a constant $c > 0$ such that*

$$\mathbb{P}\left[\|X\|_2 \geqslant t\right] \leqslant c \cdot \mathbb{P}\left[\|Z\|_2 \geqslant \frac{t - c_1}{c_2}\right]$$

for $\forall t > c_1$ where $Z \sim \mathcal{N}(0, I_d)$ is a standard Gaussian. Define truncated random variable X_{R_0} as:

$$X_{R_0} = \begin{cases} X & \textit{If } \|X\|_2 \leqslant R_0 \\ 0 & \textit{If } \|X\|_2 > R_0 \end{cases}.$$

Then, the distributional distance X and X_{R_0} is exponentially small with regard to R_0 in both Wasserstein and Total Variation metrics:

$$\mathrm{TV}(X, X_{R_0}) \lesssim \exp\left(-\frac{(R_0 - c_1)^2}{20c_2^2}\right),\ W_1(X, X_{R_0}) \lesssim \sqrt{d}c_3^d \cdot \exp\left(-\frac{(R_0 - c_1)^2}{40c_2^2}\right)$$

holds for $\forall R_0 > c_1 + \sqrt{2d} \cdot c_2$ where c_3 is a constant only dependent on c_1, c_2.

Proof of Lemma B.3. Denote $p(x)$ and $p_{R_0}(x)$ as the density function of X and X_{R_0}. Then, it is obvious that $p(x) = p_{R_0}(x)$ for $\forall 0 < \|x\|_2 \leqslant R_0$ and $p(x) \geqslant p_{R_0}(x) = 0$ for $\forall \|x\|_2 > R_0$. Another fact is that $p(x) < p_{R_0}(x)$ for $x = 0$. Therefore, the TV-distance between $p(x)$ and $p_{R_0}(x)$ can be simply expressed as:

$$\mathrm{TV}(X, X_{R_0}) = \frac{1}{2}\int |p(x) - p_{R_0}(x)|\mathrm{d}x = \int_{\|x\| \geqslant R_0} p(x)\mathrm{d}x = \mathbb{P}\left[\|X\|_2 \geqslant R_0\right].$$

Since we know that X has a (c_1, c_2)-Gaussian tail, so:

$$\mathbb{P}\left[\|X\|_2 \geqslant R_0\right] \leqslant c \cdot \mathbb{P}\left[\|Z\|_2 \geqslant \frac{(R_0 - c_1)_+}{c_2}\right] = c \cdot \mathbb{P}\left[\|Z\|_2^2 \geqslant \frac{(R_0 - c_1)_+^2}{c_2^2}\right].$$

$\|Z\|_2^2$ follows the χ_d^2 distribution, so for $\forall R_0 > c_1 + c_2 \cdot \sqrt{2d}$, its tail bound

$$\mathrm{TV}(X, X_{R_0}) \lesssim \mathbb{P}\left[\|Z\|_2^2 \geqslant \frac{(R_0 - c_1)_+^2}{c_2^2}\right] \leqslant \exp\left(-\frac{(R_0 - c_1)^2}{20c_2^2}\right).$$

For the Wasserstein-1 distance, we have the following formulation

$$\begin{aligned} W_1(X, X_{R_0}) &= \sup_{\substack{\mathrm{Lip}(f) \leqslant 1 \\ f(0)=0}} \int f(x) \cdot (p(x) - P_{R_0}(x))\,\mathrm{d}x \leqslant \int |f(x)| \cdot |p(x) - P_{R_0}(x)|\mathrm{d}x \\ &\leqslant \int \|x\|_2 \cdot |p(x) - P_{R_0}(x)|\mathrm{d}x = \int_{\|x\| > R_0} \|x\|_2 \cdot p(x)\mathrm{d}x \\ &= \mathbb{E}_x \|x\|_2 \cdot \mathbb{I}[\|x\|_2 > R_0] \leqslant \sqrt{\mathbb{E}\|x\|^2} \cdot \sqrt{\mathbb{P}[\|x\|_2 > R_0]} \lesssim \sqrt{dc_3^d} \cdot \exp\left(-\frac{(R_0 - c_1)^2}{40c_2^2}\right) \end{aligned}$$

where c_3 is a constant only related to c_1, c_2. □

According to Assumption 4.1, we know that the initial distribution p_{data} has a (α_1, α_2)-Gaussian tail. As we move forward, we propose the following properties of the Gaussian tail.

Proposition B.1. *Suppose random variable X has a (c_1, c_2)-Gaussian tail, then the following conclusions hold:*

- *For any positive constant $c_3 \geqslant c_1$ and $c_4 \geqslant c_2$, it also holds that X has a (c_3, c_4)-Gaussian tail.*
- *For positive constants $a > 0$, random variable $aX + b$ has a $(ac_1 + \|b\|_2, ac_2)$-Gaussian tail.*
- *For a-Lipschitz function F with $\|F(0)\|_2 \leqslant b$, then the random variable $F(X)$ has a $(ac_1 + b, ac_2)$-Gaussian tail.*
- *For a standard Gaussian variable $Y \sim \mathcal{N}(0, I)$, random variable $aX + bY$ has a $(ac_1, b + ac_2)$-Gaussian tail.*

Proof of Proposition B.1. According to the definition of Gaussian tail, the first statement is trivial. For $X' := aX + b$ and $\forall t > ac_1 + \|b\|_2$, we have:

$$\mathbb{P}[\|X'\| \geqslant t] \leqslant \mathbb{P}[a\|X\| \geqslant t - \|b\|] = c \cdot \mathbb{P}\left[\|X\| \geqslant \frac{t - \|b\|}{a}\right]$$

Since X has a (c_1, c_2)-Gaussian tail and $\frac{t-\|b\|}{a} \geqslant c_1$, it holds that:

$$\mathbb{P}\left[\|X\| \geqslant \frac{t-\|b\|}{a}\right] \leqslant \mathbb{P}\left[\|Z\| \geqslant \frac{1}{c_2} \cdot \left(\frac{t-\|b\|}{a} - c_1\right)\right] = \mathbb{P}\left[\|Z\| \geqslant \frac{t-\|b\|-ac_1}{ac_2}\right],$$

which comes to our second statement. For the third statement, we can simply use the result of the second statement since:

$$\|F(X)\|_2 \leqslant \|F(0)\|_2 + a\|X\|_2.$$

In fact, Statement 2 is a special case of Statement 3. For Statement 4, it holds that for $\forall \lambda \in (0,1)$:

$$\begin{aligned}\mathbb{P}[\|aX+bY\|_2 \geqslant t] &\leqslant \mathbb{P}\left[\|X\|_2 \geqslant \frac{\lambda t}{a}\right] + \mathbb{P}\left[\|Y\|_2 \geqslant \frac{(1-\lambda)t}{b}\right] \\ &\leqslant c \cdot \mathbb{P}\left[\|Z\|_2 \geqslant \frac{\lambda t - ac_1}{ac_2}\right] + \mathbb{P}\left[\|Y\|_2 \geqslant \frac{(1-\lambda)t}{b}\right].\end{aligned}$$

Let

$$\frac{\lambda t - ac_1}{ac_2} = \frac{(1-\lambda)t}{b} = \frac{t-ac_1}{b+ac_2},$$

then:

$$\mathbb{P}[\|aX+bY\|_2 \geqslant t] \leqslant (c+1) \cdot \mathbb{P}\left[\|Z\|_2 \geqslant \frac{t-ac_1}{b+ac_2}\right],$$

which means that $aX + bY$ has a $(ac_1, b+ac_2)$-Gaussian tail. □

Now, for two distributions with Gaussian tail, we show in the next lemma how to upper bound their Wasserstein distance with their total variation distance.

Lemma B.4. *For constants $c_1, c_2, d_1, d_2 > 0$, for a random variable X with (c_1, c_2)-Gaussian tail and another random variable Y with (d_1, d_2)-Gaussian tail, we can conclude that:*

$$W_1(X,Y) \leqslant C\sqrt{d} \cdot \mathrm{TV}(X,Y) \leqslant C\sqrt{d} \cdot \sqrt{\mathrm{KL}(X,Y)}$$

where C is a constant only dependent on c_1, c_2, d_1, d_2.

Proof of Lemma B.4. Denote X_{R_0}, Y_{R_0} as the truncated distributions of X, Y, then: X_{R_0}, Y_{R_0} has support set $\Omega = \{\mathbf{x} : \|\mathbf{x}\|_2 \leqslant R_0\}$. Therefore,

$$\begin{aligned}W_1(X_{R_0}, Y_{R_0}) &= \sup_{\substack{\mathrm{Lip}(f)\leqslant 1 \\ f(0)=0}} \int f(\mathbf{x})(p_{R_0}(\mathbf{x}) - q_{R_0}(\mathbf{x}))\mathrm{d}\mathbf{x} \leqslant \int \|\mathbf{x}\|_2 \cdot |p_{R_0}(\mathbf{x}) - q_{R_0}(\mathbf{x})|\mathrm{d}\mathbf{x} \\ &\leqslant R_0 \cdot \int |p_{R_0}(\mathbf{x}) - q_{R_0}(\mathbf{x})|\mathrm{d}\mathbf{x} \leqslant 2R_0 \cdot \mathrm{TV}(X_{R_0}, Y_{R_0}).\end{aligned}$$

Next, we have:

$$\begin{aligned}W_1(X,Y) &\leqslant W_1(X, X_{R_0}) + W_1(Y, Y_{R_0}) + W_1(X_{R_0}, Y_{R_0}) \\ &\leqslant \sqrt{d}c_3^d \cdot \exp\left(-\frac{(R_0-c_1)^2}{40c_2^2}\right) + \sqrt{d}d_3^d \cdot \exp\left(-\frac{(R_0-d_1)^2}{40d_2^2}\right) + 2R_0 \cdot \mathrm{TV}(X_{R_0}, Y_{R_0}) \\ &\leqslant \sqrt{d}e_3^d \cdot \exp\left(-\frac{(R_0-e_1)^2}{40e_2^2}\right) + 2R_0 \cdot (\mathrm{TV}(X,Y) + \mathrm{TV}(X, X_{R_0}) + \mathrm{TV}(Y, Y_{R_0})) \\ &\leqslant \sqrt{d}e_3^d \cdot \exp\left(-\frac{(R_0-e_1)^2}{40e_2^2}\right) + 2R_0 \cdot \exp\left(-\frac{(R_0-e_1)^2}{20e_2^2}\right) + 2R_0 \cdot \mathrm{TV}(X,Y).\end{aligned}$$

where $e_i := \max(c_i, d_i)$ for $i = 1, 2, 3$. Let $R_0 = C\sqrt{d}$ for a sufficiently large constant C, we can conclude that,

$$W_1(X,Y) \lesssim \sqrt{d} \cdot \mathrm{TV}(X,Y)$$

which comes to our lemma. □

C Proofs in Section 5

C.1 Approximation Error for Score Approximation

Lemma C.1. *Define the l-layer ReLU network class $\mathrm{NN}(l, M, J, K, \kappa, \gamma, \gamma_t)$ as follows:*

$$\begin{aligned}
&\mathrm{NN}(l, M, J, K, \kappa, \gamma, \gamma_t) = \\
&\Big\{ s(\mathbf{z}, t) = W_l \sigma(\dots \sigma(W_1[\mathbf{z}^\top, t]^\top) \dots) + b_l \ \ | \\
&\quad \textit{Network width is bounded by } M; \ \sup_{\mathbf{z}, t} \|f(\mathbf{z}, t)\|_2 \leqslant K; \\
&\quad \max_i \max(\|b_i\|_\infty, \|W_i\|_\infty) \leqslant \kappa; \ \sum_{i=1}^{l} (\|W_i\|_0 + \|b_i\|_0) \leqslant J; \\
&\quad \|s(\mathbf{z}, t) - s(\mathbf{z}', t)\|_2 \leqslant \gamma \|\mathbf{z} - \mathbf{z}'\|_2 \textit{ holds for } \forall \mathbf{z}, \mathbf{z}', t; \\
&\quad \|s(\mathbf{z}, t) - s(\mathbf{z}, t')\|_2 \leqslant \gamma_t |t - t'| \textit{ holds for } \forall \mathbf{z}, t, t' \Big\}.
\end{aligned}$$

As we see, all the neural networks in this class has bounded function value, bounded weights, bounded width and Lipschitz continuity. Given an approximation error $\delta > 0$, we choose the network hyperparameter as:

$$\begin{gathered}
l = \mathcal{O}(d + \log(1/\delta)), \ K = \mathcal{O}(2d^2 \log(d/\varepsilon\delta)), \ \gamma = 10d(1 + L), \ \gamma_t = 10\tau, \\
M = \mathcal{O}\left((1 + L)^d T \tau d^{d/2+1} \delta^{-(d+1)} \log^{d/2}(d/\varepsilon\delta)\right), \\
J = \mathcal{O}\left((1 + L)^d T \tau d^{d/2+1} \delta^{-(d+1)} \log^{d/2}(d/\varepsilon\delta)(d + \log(1/\delta))\right), \\
\kappa = \mathcal{O}\left(\max\left(2(1 + L)\sqrt{d \log(d/\varepsilon\delta)}, T\tau\right)\right)
\end{gathered}$$

where δ is chosen as $\delta = n^{-\frac{1-\tau(n)}{d+5}}$ for $\tau(n) = \frac{d \log\log n}{\log n}$ and

$$\tau := \sup_t \sup_{\|\mathbf{z}\|_\infty \leqslant \sqrt{d \log(d/\varepsilon\delta)}} \left\| \frac{\partial}{\partial t} \left[\sigma(t)^2 \nabla \log p_t(\mathbf{z})\right] \right\|_2 .$$

After choosing There exists $s_\phi \in \mathrm{NN}$ such that with probability at least $1 - \frac{1}{n}$, it holds that

$$\frac{1}{t_b - t_a} \int_{t_a}^{t_b} \|s_\phi(\cdot, t) - \nabla \log p_t(\cdot)\|^2_{L^2(\mathbf{X}_t)} \, \mathrm{d}t = \widetilde{O}\left(\frac{1}{\varepsilon} n^{-\frac{2}{d+5}}\right) \quad \forall \varepsilon \leqslant t_a < t_b \leqslant T.$$

C.2 Proof of Corollary 5.1

For any given $0 \leqslant k < N$, denote L_k to be the Lipschitz constant of $f^*_\theta(\cdot, t)$ when $t \in [t_k, t_{k+1}]$. If we treat $\mathbf{x}^\phi$ as a function over $\mathbf{x}$, its Lipschitz constant is no larger than

$$1 + \overline{\beta}(1 + L_{\text{score}})\Delta t/2 \leqslant 1 + Cd\overline{\beta}\Delta t$$

where $C = 10(1 + L)$ is a pure constant. This is also the upper bound of L_1. Here, we use the result in Lemma 5.1 that $L_{\text{score}} = \mathcal{O}(10d(1 + L))$. Therefore:

$$L_{k+1} \leqslant (1 + Cd\overline{\beta}\Delta t)L_k$$

holds according to the recursive formulation of $f_{\theta^*}(\cdot, \cdot)$, which leads to the conclusion that, the Lipschitz constant of $f_{\theta^*}(\cdot, t)$ is no larger than:

$$(1 + Cd\overline{\beta}\Delta t)^N = (1 + Cd\overline{\beta}\Delta t)^{T/\Delta t} \leqslant \exp(Cd\overline{\beta}T),$$

which proves the conclusion.

C.3 Proof of Lemma 5.2

As described in the lemma, we recall that

$$\begin{aligned}
I_1 &:= \sum_{k=1}^{N'} W_1\left(f_{\widehat{\theta}}(\cdot, \tau_{k-1})_\sharp \mathbf{X}_{\tau_{k-1}}, f_{\widehat{\theta}}(\cdot, \tau_{k-1})_\sharp \widehat{\mathbf{X}}^{\phi, M}_{\tau_{k-1}}\right), \\
I_2 &:= \sum_{k=1}^{N'} W_1\left(f_{\theta^*}(\cdot, \tau_{k-1})_\sharp \mathbf{X}_{\tau_{k-1}}, f_{\theta^*}(\cdot, \tau_{k-1})_\sharp \widehat{\mathbf{X}}^{\phi, M}_{\tau_{k-1}}\right), \\
I_3 &:= \sum_{k=1}^{N'} \left[W_1\left(f_{\widehat{\theta}}(\cdot, \tau_k)_\sharp \mathbf{X}_{\tau_k}, f_{\widehat{\theta}}(\cdot, \tau_{k-1})_\sharp \widehat{\mathbf{X}}^{\phi, M}_{\tau_{k-1}}\right) - W_1\left(f_{\widehat{\theta}}(\cdot, \tau_k)_\sharp \mathcal{X}_{\tau_k}, f_{\widehat{\theta}}(\cdot, \tau_{k-1})_\sharp \widehat{\mathcal{X}}^{\phi, M}_{\tau_{k-1}}\right)\right] \\
I_4 &:= \sum_{k=1}^{N'} \left[W_1\left(f_{\theta^*}(\cdot, \tau_k)_\sharp \mathcal{X}_{\tau_k}, f_{\theta^*}(\cdot, \tau_{k-1})_\sharp \widehat{\mathcal{X}}^{\phi, M}_{\tau_{k-1}}\right) - W_1\left(f_{\theta^*}(\cdot, \tau_k)_\sharp \mathbf{X}_{\tau_k}, f_{\theta^*}(\cdot, \tau_{k-1})_\sharp \widehat{\mathbf{X}}^{\phi, M}_{\tau_{k-1}}\right)\right].
\end{aligned}$$

First, from the optimality condition (18) and the structure of f_θ^* (19), we have:

$$\begin{aligned}
&\sum_{k=1}^{N'} W_1\left(f_{\widehat{\theta}}(\cdot,\tau_k)_\sharp \mathcal{X}_{\tau_k}, f_{\widehat{\theta}}(\cdot,\tau_{k-1})_\sharp \widehat{\mathcal{X}}^{\phi,M}_{\tau_{k-1}}\right) \leqslant \sum_{k=1}^{N'} W_1\left(f_{\theta^*}(\cdot,\tau_k)_\sharp \mathcal{X}_{\tau_k}, f_{\theta^*}(\cdot,\tau_{k-1})_\sharp \widehat{\mathcal{X}}^{\phi,M}_{\tau_{k-1}}\right)\\
&\leqslant \sum_{k=1}^{N'}\left[W_1\left(f_{\theta^*}(\cdot,\tau_k)_\sharp \mathcal{X}_{\tau_k}, f_{\theta^*}(\cdot,\tau_{k-1})_\sharp \widehat{\mathcal{X}}^{\phi,M}_{\tau_{k-1}}\right) - W_1\left(f_{\theta^*}(\cdot,\tau_k)_\sharp \mathbf{X}_{\tau_k}, f_{\theta^*}(\cdot,\tau_{k-1})_\sharp \widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}\right)\right]\\
&+\sum_{k=1}^{N'} W_1\left(f_{\theta^*}(\cdot,\tau_k)_\sharp \mathbf{X}_{\tau_k}, f_{\theta^*}(\cdot,\tau_{k-1})_\sharp \widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}\right) = I_4.
\end{aligned}$$

Then, we can immediately conclude that:

$$\sum_{k=1}^{N'} W_1\left(f_{\widehat{\theta}}(\cdot,\tau_k)_\sharp \mathbf{X}_{\tau_k}, f_{\widehat{\theta}}(\cdot,\tau_{k-1})_\sharp \widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}\right) \leqslant I_3 + I_4. \tag{25}$$

Again by using Equation (19), we know that for $\forall k \in [N']$:

$$\begin{aligned}
&W_1\left(f_{\widehat{\theta}}(\cdot,\tau_k)_\sharp \mathbf{X}_{\tau_k}, f_{\theta^*}(\cdot,\tau_k)_\sharp \mathbf{X}_{\tau_k}\right) = W_1\left(f_{\widehat{\theta}}(\cdot,\tau_k)_\sharp \mathbf{X}_{\tau_k}, f_{\theta^*}(\cdot,\tau_{k-1})_\sharp \widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}\right)\\
&\leqslant W_1\left(f_{\widehat{\theta}}(\cdot,\tau_k)_\sharp \mathbf{X}_{\tau_k}, f_{\widehat{\theta}}(\cdot,\tau_{k-1})_\sharp \widehat{\mathbf{X}}^{\phi,M}_{\tau_k}\right) + W_1\left(f_{\widehat{\theta}}(\cdot,\tau_{k-1})_\sharp \mathbf{X}_{\tau_{k-1}}, f_{\widehat{\theta}}(\cdot,\tau_{k-1})_\sharp \widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}\right)\\
&+W_1\left(f_{\widehat{\theta}}(\cdot,\tau_{k-1})_\sharp \mathbf{X}_{\tau_{k-1}}, f_{\theta^*}(\cdot,\tau_{k-1})_\sharp \mathbf{X}_{\tau_{k-1}}\right) + W_1\left(f_{\theta^*}(\cdot,\tau_{k-1})_\sharp \mathbf{X}_{\tau_{k-1}}, f_{\theta^*}(\cdot,\tau_{k-1})_\sharp \widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}\right)
\end{aligned} \tag{26}$$

Then, after summing over $k = 1, 2, \dots, N'$ and telescoping, we have:

$$\begin{aligned}
&W_1\left(f_{\widehat{\theta}}(\cdot,T)_\sharp \mathbf{X}_T, f_{\theta^*}(\cdot,T)_\sharp \mathbf{X}_T\right) \leqslant \sum_{k=1}^{N'} W_1\left(f_{\widehat{\theta}}(\cdot,\tau_k)_\sharp \mathbf{X}_{\tau_k}, f_{\widehat{\theta}}(\cdot,\tau_{k-1})_\sharp \widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}\right)\\
&+\sum_{k=1}^{N'} W_1\left(f_{\widehat{\theta}}(\mathbf{X}_{\tau_{k-1}},\tau_{k-1}), f_{\widehat{\theta}}(\cdot,\tau_{k-1})_\sharp \widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}\right) + \sum_{k=1}^{N'} W_1\left(f_{\theta^*}(\cdot,\tau_{k-1})_\sharp \mathbf{X}_{\tau_{k-1}}, f_{\theta^*}(\cdot,\tau_{k-1})_\sharp \widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}\right)\\
&=\sum_{k=1}^{N'} W_1\left(f_{\widehat{\theta}}(\cdot,\tau_k)_\sharp \mathbf{X}_{\tau_k}, f_{\widehat{\theta}}(\cdot,\tau_{k-1})_\sharp \widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}\right) + I_1 + I_2 \leqslant I_1 + I_2 + I_3 + I_4.
\end{aligned}$$

Here, we apply Equation (25) to the last line, and finally we come to our conclusion.

C.4 Proof of Lemma 5.3

According to the optimization constraint, we know that $f_{\widehat{\theta}}, f_{\theta^*} \in \mathrm{Lip}(R)$. Therefore, we can combine these two terms I_1, I_2 and see how to upper bound

$$J := \sup_{f_\theta \in \mathrm{Lip}(R)} \sum_{k=1}^{N'} W_1\left(f_\theta(\cdot,\tau_{k-1})_\sharp \mathbf{X}_{\tau_{k-1}}, f_\theta(\cdot,\tau_{k-1})_\sharp \widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}\right).$$

Notice that, $\mathbf{X}_{\tau_{k-1}}, \mathbf{X}_{\tau_k}$ are sampled from the forward process, which means

$$\mathbf{X}_t = (m(t)\cdot p_{\mathrm{data}}) \star \mathcal{N}(0,\sigma(t)^2) \quad \forall t \in [\varepsilon, T]$$

where $m(t) = \exp\left(-\int_0^t \beta(s)\mathrm{d}s\right), \sigma(t)^2 = 1 - m(t)^2$ and $\star$ denotes the convolution between two distributions. Also, $\widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}$ is sampled from multi-step discretization of backward probability ODE flow (Equation (5)), starting from $\mathbf{X}_{\tau_k}$. Lemma B.1 provides us an upper bound for the KL-divergence between $\mathbf{X}_{\tau_{k-1}}$ and $\widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}$. Since $f_\theta \in \mathrm{Lip}(R)$, $f_\theta(\cdot,\tau_{k-1})$ is an R-Lipschitz function, so it holds that:

$$W_1\left(f_\theta(\cdot,\tau_{k-1})_\sharp \mathbf{X}_{\tau_{k-1}}, f_\theta(\cdot,\tau_{k-1})_\sharp \widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}\right) \leqslant R \cdot W_1\left(\mathbf{X}_{\tau_{k-1}}, \widehat{\mathbf{X}}^{\phi}_{\tau_{k-1}}\right).$$

In order to upper bound $W_1\left(\mathbf{X}_{\tau_{k-1}}, \widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}\right)$ with the KL divergence $\mathrm{KL}\left(\mathbf{X}_{\tau_{k-1}}, \widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}\right)$, we need to apply Lemma B.4. In the following part, we prove that the random variable $\mathbf{x}^{\phi,M}_{\tau_{k-1}}$ has Gaussian tail, just like $\mathbf{x}_{\tau_k}$. For any integer $k \in [1, N]$, it holds that:

$$\left\|\widehat{\mathbf{x}}^{\phi}_{t_{k-1}}\right\|_2 = \left\|\mathbf{x}_{t_k} + \left(\frac{\beta(t_k)}{2}\mathbf{x}_{t_k} + \frac{\beta(t_k)}{2}s_\phi(\mathbf{x}_{t_k}, t)\right)\cdot \Delta t\right\|_2$$
$$< \left(1 + \frac{\overline{\beta}\Delta t}{2}\right)\cdot \|\mathbf{x}_{t_k}\|_2 + \frac{\overline{\beta}U_{\text{score}}\cdot \Delta t}{2} < \left(1 + \frac{\Delta t}{U_{\text{score}}}\right)\|\mathbf{x}_{t_k}\|_2 + \Delta t$$

when $\overline{\beta} < 2/U_{\text{score}}$. After iterating this inequality M times, we have:

$$\left\|\widehat{\mathbf{x}}^{\phi,M}_{\tau_{k-1}}\right\|_2 \leqslant \left(1 + \frac{\Delta t}{U_{\text{score}}}\right)^M \|\mathbf{x}_{\tau_k}\|_2 + \Delta t \cdot \left(1 + \left(1 + \frac{\Delta t}{U_{\text{score}}}\right) + \ldots + \left(1 + \frac{\Delta t}{U_{\text{score}}}\right)^{M-1}\right).$$

Under the condition that $N' \gg T$, we have:

$$\left(1 + \frac{\Delta t}{U_{\text{score}}}\right)^M < (1 + \Delta t)^{T/N'\Delta t} < \exp(T/N') < 2,$$

which leads to:

$$\Delta t \cdot \left(1 + \left(1 + \frac{\Delta t}{U_{\text{score}}}\right) + \ldots + \left(1 + \frac{\Delta t}{U_{\text{score}}}\right)^{M-1}\right) \leqslant \Delta t \cdot 2M = \frac{2T}{N'} < 2.$$

Therefore, we have $\left\|\widehat{\mathbf{x}}^{\phi,M}_{t_{k-1}}\right\|_2 \leqslant 2\|\mathbf{x}_{t_k}\|_2 + 2$, which means both $\mathbf{X}_{\tau_k}$ and $\mathbf{X}^{\phi,M}_{\tau_{k-1}}$ have Gaussian tail for all $k \in [1, N']$. By applying Lemma B.4, we conclude that:

$$\begin{aligned} I_1 + I_2 &\leqslant 2R \cdot \sum_{k=1}^{N'} W_1\left(\mathbf{X}_{\tau_{k-1}}, \widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}\right) \lesssim 2R\sqrt{d} \cdot \sum_{k=1}^{N'} \sqrt{\mathrm{KL}\left(\mathbf{X}_{\tau_{k-1}}, \widehat{\mathbf{X}}^{\phi,M}_{\tau_{k-1}}\right)} \\ &\lesssim 2R\sqrt{d}N' \cdot \sqrt{\overline{\beta}^2\left(L^2 d\Delta t \cdot M\Delta t + \frac{1}{\varepsilon}n^{-\frac{2}{d+5}} \cdot M\Delta t\right)} \\ &\lesssim RdL\overline{\beta} \cdot \frac{T}{\sqrt{M}} + R\overline{\beta}\sqrt{d}n^{-\frac{1}{d+5}} \cdot \sqrt{\frac{N'T}{\varepsilon}}. \end{aligned} \tag{27}$$

After arranging these terms, we come to our conclusion.

C.5 Proof of Lemma 5.4

For these two loss terms, they can be treated as the gap between empirical and population Wasserstein distances. Notice that, for any two distributions p, q, denote $\widehat{p}, \widehat{q}$ as their empirical version, then it holds that:

$$|W_1(p, q) - W_1(\widehat{p}, \widehat{q})| \leqslant W_1(p, \widehat{p}) + W_1(q, \widehat{q}). \tag{28}$$

Notice that $f_{\widehat{\theta}}(\cdot, \tau_k), f_{\theta^*}(\cdot, \tau_k)$ are Lipschitz-R continuous function for all $k \in [1, N']$. Also, since:

$$G(\mathbf{x}, t_k; \phi) = \mathbf{x} + \left(\frac{\beta(t_k)}{2}\mathbf{x} + \frac{\beta(t_k)}{2}s_\phi(\mathbf{x}, t_k)\right)\cdot \Delta t,$$

which is Lipschitz continuous with regard to $\mathbf{x}_{t_k}$ with Lipschitz constant

$$L_1 = 1 + \overline{\beta}(1 + L_{\text{score}})\Delta t/2 < 1 + \Delta t$$

since $\overline{\beta} < 2/(1 + L_{\text{score}})$. After iterating M times, we know that:

$$G_{(M)}(\cdot, \tau_k; \phi) = G(\cdot, t_{(k-1)M+1}; \phi) \circ \ldots \circ G(\cdot, t_{kM}; \phi)$$

is Lipschitz continuous with constant $(1+\Delta t)^M < \exp(T/N') < 2$. Therefore, $f_{\widehat{\theta}}(\widehat{\mathbf{x}}^{\phi,M}_{\tau_{k-1}}, \tau_{k-1})$ and $f_{\theta^*}(\widehat{\mathbf{x}}^{\phi,M}_{\tau_{k-1}}, \tau_{k-1})$ are $2R$-Lipschitz continuous function with regard to $\mathbf{x}_{\tau_k}$. According to Inequality (28), we have:

$$
\begin{aligned}
|I_3| &\leqslant \sum_{k=1}^{N'} \left[W_1\left(f_{\widehat{\theta}}(\cdot,\tau_k)_\sharp \mathbf{X}_{\tau_k}, f_{\widehat{\theta}}(\cdot,t_k)_\sharp \mathcal{X}_{\tau_k}\right) + W_1\left(f_{\widehat{\theta}}(\cdot,\tau_{k-1})_\sharp \widehat{\mathbf{X}}^{\phi,M}_{\tau_k}, f_{\widehat{\theta}}(\cdot,\tau_{k-1})_\sharp \widehat{\mathcal{X}}^{\phi,M}_{\tau_k}\right)\right] \\
&\leqslant \sum_{k=1}^{N'} \left[R\cdot W_1\left(\mathbf{X}_{\tau_k},\mathcal{X}_{\tau_k}\right) + R\cdot W_1\left(G_{(M)}(\cdot,\tau_k;\phi)_\sharp \mathbf{X}_{\tau_k}, G_{(M)}(\cdot,\tau_k;\phi)_\sharp \mathcal{X}_{\tau_k}\right)\right] \\
&\leqslant \sum_{k=1}^{N'} \left[R\cdot W_1\left(\mathbf{X}_{\tau_k},\mathcal{X}_{\tau_k}\right) + 2R\cdot W_1\left(\mathbf{X}_{\tau_k},\mathcal{X}_{\tau_k}\right)\right] = 3R\cdot \sum_{k=1}^{N'} W_1\left(\mathbf{X}_{\tau_k},\mathcal{X}_{\tau_k}\right). \\
|I_4| &\leqslant \sum_{k=1}^{N'} \left[W_1\left(f_{\theta^*}(\cdot,\tau_k)_\sharp \mathbf{X}_{\tau_k}, f_{\theta^*}(\cdot,t_k)_\sharp \mathcal{X}_{\tau_k}\right) + W_1\left(f_{\theta^*}(\cdot,\tau_{k-1})_\sharp \widehat{\mathbf{X}}^{\phi,M}_{\tau_k}, f_{\theta^*}(\cdot,\tau_{k-1})_\sharp \widehat{\mathcal{X}}^{\phi,M}_{\tau_k}\right)\right] \\
&\leqslant \sum_{k=1}^{N'} \left[R\cdot W_1\left(\mathbf{X}_{\tau_k},\mathcal{X}_{\tau_k}\right) + R\cdot W_1\left(G_{(M)}(\cdot,\tau_k;\phi)_\sharp \mathbf{X}_{\tau_k}, G_{(M)}(\cdot,\tau_k;\phi)_\sharp \mathcal{X}_{\tau_k}\right)\right] \\
&\leqslant \sum_{k=1}^{N'} \left[R\cdot W_1\left(\mathbf{X}_{\tau_k},\mathcal{X}_{\tau_k}\right) + 2R\cdot W_1\left(\mathbf{X}_{\tau_k},\mathcal{X}_{\tau_k}\right)\right] = 3R\cdot \sum_{k=1}^{N'} W_1\left(\mathbf{X}_{\tau_k},\mathcal{X}_{\tau_k}\right).
\end{aligned}
\tag{29}
$$

Here, $\mathcal{X}_{\tau_k}$ is the empirical version of distribution $\mathbf{X}_{\tau_k}$, which means:

$$
\mathbf{X}_{\tau_k} = (m(\tau_k)\cdot p_{\text{data}}) \star \mathcal{N}(0,\sigma(\tau_k)^2),\ \mathcal{X}_{\tau_k} = (m(\tau_k)\cdot \widehat{p_{\text{data}}}) \star \mathcal{N}(0,\sigma(\tau_k)^2)
$$

where $\widehat{p_{\text{data}}}$ is a uniform distribution taken over the n i.i.d samples from p_{data}. In order to upper bound $|I_3|, |I_4|$, we only need to control $W_1\left(\mathbf{X}_{\tau_k},\mathcal{X}_{\tau_k}\right)$. The following lemma upper bounds $W_1\left(\mathbf{X}_{\tau_k},\mathcal{X}_{\tau_k}\right)$ with $W_1\left(p_{\text{data}},\widehat{p_{\text{data}}}\right)$ for each $k=1,2,\dots,N'$.

Lemma C.2. *For each $k=1,2,\dots,N'$, it holds that:*

$$
W_1\left(\mathbf{X}_{\tau_k},\mathcal{X}_{\tau_k}\right) \leqslant m(\tau_k)\cdot W_1\left(p_{\text{data}},\ \widehat{p_{\text{data}}}\right).
$$

Proof. By the dual formulation of Wasserstein distance, it holds that for $\forall t\in[0,T]$:

$$
W_1(\mathbf{X}_t,\mathcal{X}_t) = \sup_{\text{Lip}(F)\leqslant 1}\left(\mathbb{E}_{\mathbf{x}\sim\mathbf{X}_t}F(\mathbf{x}) - \mathbb{E}_{\mathbf{x}\sim\mathcal{X}_t}F(\widehat{\mathbf{x}})\right) = \sup_{\text{Lip}(F)\leqslant 1}\left[\mathbb{E}_{\mathbf{x}}\mathbb{E}_z F(m_t\mathbf{x}+\sigma_t z) - \mathbb{E}_{\widehat{\mathbf{x}}}\mathbb{E}_z F(m_t\widehat{\mathbf{x}}+\sigma_t z)\right]
$$

where the expectation is taken over $\mathbf{x}\sim p_{\text{data}}, \widehat{\mathbf{x}}\sim\widehat{p_{\text{data}}}$ and $z\sim\mathcal{N}(0,\boldsymbol{I})$. Notice that, for the following mapping

$$
G[F](\mathbf{x}) := \mathbb{E}_z F(m(t)\mathbf{x}+\sigma(t)z),
$$

it holds that for any function F with Lipschitz constant 1,

$$
|G[F](\mathbf{x}) - G[F](\mathbf{x}')| \leqslant \mathbb{E}_z\left|F(m(t)\mathbf{x}+\sigma(t)z) - F(m(t)\mathbf{x}'+\sigma(t)z)\right| \leqslant \mathbb{E}_z\left[m(t)\cdot|\mathbf{x}-\mathbf{x}'|\right] = m(t)\cdot|\mathbf{x}-\mathbf{x}'|.
$$

Therefore, we know that $G[F]$ is $m(t)$-Lipschitz continuous, which leads to

$$
\begin{aligned}
W_1(\mathbf{X}_t,\widehat{\mathbf{X}_t}) &= \sup_{\text{Lip}(F)\leqslant 1}\left[\mathbb{E}_{\mathbf{x}}G[F](\mathbf{x}) - \mathbb{E}_{\widehat{\mathbf{x}}}G[F](\widehat{\mathbf{x}})\right] \leqslant \sup_{\text{Lip}(G)\leqslant m_t}\left[\mathbb{E}_{\mathbf{x}}G(\mathbf{x}) - \mathbb{E}_{\widehat{\mathbf{x}}}G(\widehat{\mathbf{x}})\right] \\
&= m(t)\cdot\sup_{\text{Lip}(G)\leqslant 1}\left[\mathbb{E}_{\mathbf{x}}G(\mathbf{x}) - \mathbb{E}_{\widehat{\mathbf{x}}}G(\widehat{\mathbf{x}})\right] = m(t)\cdot W_1(p_{\text{data}},\ \widehat{p_{\text{data}}}).
\end{aligned}
$$

It comes to our conclusion. □

After that, by combining Equation (29) and Lemma C.2, we have:

$$
|I_3| + |I_4| \leqslant 6R\left(\sum_{k=1}^{N'} m(\tau_k)\right)\cdot W_1\left(p_{\text{data}},\widehat{p_{\text{data}}}\right) < 6RN'\cdot W_1\left(p_{\text{data}},\widehat{p_{\text{data}}}\right). \tag{30}
$$

Now, our final step is to bound the gap between empirical and population Wasserstein distance of the initial distribution p_{data}. According to the statistical result (Weed & Bach, 2019), we can conclude that:

$$
\mathbb{E}W_1\left(p_{\text{data}},\widehat{p}_{\text{data}}\right) \lesssim n^{-1/d}.
$$

Here, the expectation is taken over $\widehat{p}_{\text{data}} \overset{i.i.d}{\sim} p_{\text{data}}$. Therefore, we can upper bound I_3+I_4 as:

$$
I_3 + I_4 \leqslant |I_3| + |I_4| \lesssim 6RN'\cdot n^{-1/d}. \tag{31}
$$

C.6 Proof of Lemma 5.5

From the proof of Lemma C.2, we know that

$$W_1\left(P \star \mathcal{N}(0,\sigma^2\boldsymbol{I}), Q \star \mathcal{N}(0,\sigma^2\boldsymbol{I})\right) \leqslant W_1(P,Q)$$

holds for any distribution pair (P,Q) and $\sigma > 0$. For distribution $\mathbf{X}_T$ and $\mathcal{N}(0,\boldsymbol{I})$, we have:

$$\mathbf{X}_T = (m(T)\cdot p_{\mathrm{data}}) \star \mathcal{N}(0,\sigma(T)^2) \text{ and } \mathcal{N}(0,\boldsymbol{I}) = (m(T)\cdot \mathcal{N}(0,\boldsymbol{I})) \star \mathcal{N}(0,\sigma(T)^2)$$

since $m(T)^2 + \sigma(T)^2 = 1$. Therefore:

$$W_1(\mathbf{X}_T, \mathcal{N}(0,\boldsymbol{I})) \leqslant m(T)\cdot W_1(p_{\mathrm{data}}, \mathcal{N}(0,\boldsymbol{I})).$$

According to Assumption 4.2, we know that $W_1(p_{\mathrm{data}}, \mathcal{N}(0,\boldsymbol{I}))$ is finite and furthermore $W_1(p_{\mathrm{data}}, \mathcal{N}(0,\boldsymbol{I})) \lesssim \sqrt{d}$. Besides,

$$m(T) = \exp\left(-\frac{1}{2}\int_0^T \beta(s)\mathrm{d}s\right) \leqslant \exp(-\underline{\beta}T/2).$$

To sum up, we conclude that:

$$W_1(\mathbf{X}_T, \mathcal{N}(0,\boldsymbol{I})) \lesssim \sqrt{d}\exp(-\underline{\beta}T/2).$$

C.7 Proof of Lemma 5.6

Lemma C.3 (DDPM)**.** *Under Assumption 4.1 and 4.2, when the step size $\Delta t < 1/L$, it holds that:*

$$\mathrm{KL}\left(f_{\theta^*}(\cdot,T)_\sharp \mathbf{X}_T, \mathbf{X}_\varepsilon\right) \lesssim \overline{\beta}^2 L^2 T(d\Delta t + \mathcal{M}_2^2\Delta t^2) + \overline{\beta}^2 \int_\varepsilon^T \|s_\phi(\cdot,t) - \nabla\log p_t(\cdot)\|^2_{L^2(\mathbf{X}_t)}\,\mathrm{d}t$$

Since both $\mathbf{X}_\varepsilon$ and $f_{\theta^*}(\cdot,T)_\sharp \mathbf{X}_T$ have Gaussian tail, we apply Lemma B.4 and the score integrated error (Lemma 5.1), then we conclude that:

$$W_1\left(f_{\theta^*}(\cdot,T)_\sharp \mathbf{X}_T, \mathbf{X}_\varepsilon\right) \lesssim \sqrt{d}\cdot\sqrt{\mathrm{KL}\left(f_{\theta^*}(\cdot,T)_\sharp \mathbf{X}_T, \mathbf{X}_\varepsilon\right)} \lesssim \overline{\beta}Ld\sqrt{T\Delta t} + \overline{\beta}\sqrt{d}\cdot\sqrt{\frac{T}{\varepsilon}}n^{-\frac{1}{d+5}}.$$

Finally, we just need to bound $W_1(\mathbf{X}_\varepsilon, p_{\mathrm{data}})$, which is stated in the following lemma.

Lemma C.4. *For the distributions p_{data} and $\mathbf{X}_\varepsilon = (m(\varepsilon)\cdot p_{\mathrm{data}}) \star \mathcal{N}(0,\sigma(\varepsilon)^2\boldsymbol{I})$, its Wasserstein-1 distance with p_{data} can be upper bounded as:*

$$W_1(\mathbf{X}_\varepsilon, p_{\mathrm{data}}) \lesssim \sqrt{d\overline{\beta}\varepsilon}.$$

Proof. Notice that $\mathbf{X}_\varepsilon = (m(\varepsilon)\cdot p_{\mathrm{data}}) \star \mathcal{N}(0,\sigma(\varepsilon)^2)$ where

$$m(\varepsilon) = \exp\left(-\frac{1}{2}\int_0^\varepsilon \beta(s)\mathrm{d}s\right) \geqslant \exp(-\overline{\beta}\varepsilon/2) \geqslant 1 - \overline{\beta}\varepsilon/2,$$

and $\sigma(\varepsilon)^2 = 1 - m(\varepsilon)^2 \leqslant 2(1-m(\varepsilon)) \leqslant \overline{\beta}\varepsilon$. Then, it holds that:

$$\begin{aligned}
W_1(\mathbf{X}_\varepsilon, p_{\mathrm{data}}) &\leqslant W_1\left(p_{\mathrm{data}}, m(\varepsilon)\cdot p_{\mathrm{data}}\right) + W_1\left(m(\varepsilon)\cdot p_{\mathrm{data}}, \mathbf{X}_\varepsilon\right) \\
&\leqslant \sup_{\mathrm{Lip}(f)\leqslant 1} \mathbb{E}_{\mathbf{x}\sim p_{\mathrm{data}}}\left[f(\mathbf{x}) - f(m(\varepsilon)\cdot\mathbf{x})\right] + W_1(\delta_{\{0\}}, \mathcal{N}(0,\sigma(\varepsilon)^2)) \\
&\leqslant (1-m(\varepsilon))\cdot\mathbb{E}_{\mathbf{x}\sim p_{\mathrm{data}}}\|\mathbf{x}\|_2 + \sigma(\varepsilon)\cdot\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(0,\boldsymbol{I})}\|z\|_2 \\
&\leqslant (1-m(\varepsilon))\cdot\mathcal{M}_2 + \sigma(\varepsilon)\cdot\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(0,\boldsymbol{I})}\|\mathbf{z}\|_2 \\
&\leqslant \overline{\beta}\varepsilon/2\cdot\mathcal{M}_2 + \sqrt{\overline{\beta}\varepsilon}\cdot\sqrt{d} \lesssim \sqrt{d\overline{\beta}\varepsilon},
\end{aligned}$$

which comes to our conclusion. □

D Proof Sketch for Consistency Isolation

Unlike the distillation case, the consistency equality we apply is based on the empirical distributions, so that θ^* satisfies

$$f_{\theta^*}(\cdot, \tau_k)_\sharp \mathcal{X}_{\tau_k} \overset{\text{law}}{=} f_{\theta^*}(\cdot, \tau_{k-1})_\sharp \widehat{\mathcal{X}}^M_{\tau_{k-1}} \quad \forall k \in [N']$$

because of the definition of f_{θ^*}. Besides, according to the optimality inequality, we have:

$$\sum_{k=1}^{N'} W_1\left(f_{\widehat{\theta}}(\cdot, \tau_k)_\sharp \mathcal{X}_{\tau_k}, f_{\widehat{\theta}}(\cdot, \tau_{k-1})_\sharp \mathcal{X}_{\tau_{k-1}}\right) \leqslant \sum_{k=1}^{N'} W_1\left(f_{\theta^*}(\cdot, \tau_k)_\sharp \mathcal{X}_{\tau_k}, f_{\theta^*}(\cdot, \tau_{k-1})_\sharp \mathcal{X}_{\tau_{k-1}}\right).$$

After combining these two inequalities, we can upper bound our target function as follows:

Lemma D.1.

$$W_1\left(f_{\widehat{\theta}}(\cdot, T)_\sharp \mathcal{X}_T, f_{\theta^*}(\cdot, T)_\sharp \mathcal{X}_T\right) \leqslant 2\sum_{k=1}^{N'} W_1\left(f_{\theta^*}(\cdot, \tau_{k-1})_\sharp \widehat{\mathcal{X}}^M_{\tau_{k-1}}, f_{\theta^*}(\cdot, \tau_{k-1})_\sharp \mathcal{X}_{\tau_{k-1}}\right).$$

Proof. Notice that for $\forall k \in [N']$, we have:

$$\begin{aligned} W_1\left(f_{\widehat{\theta}}(\cdot, \tau_k)_\sharp \mathcal{X}_{\tau_k}, f_{\theta^*}(\cdot, \tau_k)_\sharp \mathcal{X}_{\tau_k}\right) &\leqslant W_1\left(f_{\widehat{\theta}}(\cdot, \tau_{k-1})_\sharp \mathcal{X}_{\tau_{k-1}}, f_{\theta^*}(\cdot, \tau_{k-1})_\sharp \mathcal{X}_{\tau_{k-1}}\right) \\ &+ W_1\left(f_{\widehat{\theta}}(\cdot, \tau_k)_\sharp \mathcal{X}_{\tau_k}, f_{\widehat{\theta}}(\cdot, \tau_{k-1})_\sharp \mathcal{X}_{\tau_{k-1}}\right) + W_1\left(f_{\theta^*}(\cdot, \tau_k)_\sharp \mathcal{X}_{\tau_k}, f_{\theta^*}(\cdot, \tau_{k-1})_\sharp \mathcal{X}_{\tau_{k-1}}\right). \end{aligned}$$

After taking summation over $k = 1, 2, \ldots, N'$, we have:

$$\begin{aligned} &W_1\left(f_{\widehat{\theta}}(\cdot, T)_\sharp \mathcal{X}_T, f_{\theta^*}(\cdot, T)_\sharp \mathcal{X}_T\right) \\ &\leqslant \sum_{k=1}^{N'} W_1\left(f_{\widehat{\theta}}(\cdot, \tau_k)_\sharp \mathcal{X}_{\tau_k}, f_{\widehat{\theta}}(\cdot, \tau_{k-1})_\sharp \mathcal{X}_{\tau_{k-1}}\right) + \sum_{k=1}^{N'} W_1\left(f_{\theta^*}(\cdot, \tau_k)_\sharp \mathcal{X}_{\tau_k}, f_{\theta^*}(\cdot, \tau_{k-1})_\sharp \mathcal{X}_{\tau_{k-1}}\right) \\ &\leqslant \sum_{k=1}^{N'} W_1\left(f_{\theta^*}(\cdot, \tau_k)_\sharp \mathcal{X}_{\tau_k}, f_{\theta^*}(\cdot, \tau_{k-1})_\sharp \mathcal{X}_{\tau_{k-1}}\right) + \sum_{k=1}^{N'} W_1\left(f_{\theta^*}(\cdot, \tau_k)_\sharp \mathcal{X}_{\tau_k}, f_{\theta^*}(\cdot, \tau_{k-1})_\sharp \mathcal{X}_{\tau_{k-1}}\right) \\ &= 2\sum_{k=1}^{N'} W_1\left(f_{\theta^*}(\cdot, \tau_{k-1})_\sharp \widehat{\mathcal{X}}^M_{\tau_{k-1}}, f_{\theta^*}(\cdot, \tau_{k-1})_\sharp \mathcal{X}_{\tau_{k-1}}\right), \end{aligned}$$

which comes to our conclusion. Here, we use the optimality inequality as well as the consistency equation. □

As we can see, the loss decomposition is much simpler than the distillation case. The only relevant term stands for the discretization error of ODE solver. Since $f_{\theta^*}(\cdot, t)$ is R-Lipschitz for any $t \in [0, 1]$, we have

$$W_1\left(f_{\theta^*}(\cdot, \tau_{k-1})_\sharp \widehat{\mathcal{X}}^M_{\tau_{k-1}}, f_{\theta^*}(\cdot, \tau_{k-1})_\sharp \mathcal{X}_{\tau_{k-1}}\right) \leqslant R \cdot W_1\left(\widehat{\mathcal{X}}^M_{\tau_{k-1}}, \mathcal{X}_{\tau_{k-1}}\right).$$

Compared with Equation (24), we do not have the score approximation error here since the score function for $\mathcal{X}_t$ has explicit formulation, which leads to

$$\mathrm{KL}\left(\widehat{\mathcal{X}}^M_{\tau_{k-1}}, \mathcal{X}_{\tau_{k-1}}\right) \lesssim \overline{\beta}^2 L_\varepsilon^2 d\Delta t \cdot M\Delta t.$$

Here, the score function $\nabla \log \widehat{p}_t(\cdot)$ is L_ε-Lipschitz continuous for $\forall t \in [\varepsilon, T]$. By using Lemma B.4 and Lemma D.1, we have:

$$W_1\left(f_{\widehat{\theta}}(\cdot, T)_\sharp \mathcal{X}_T, f_{\theta^*}(\cdot, T)_\sharp \mathcal{X}_T\right) \leqslant 2R \cdot \sum_{k=1}^{N'} W_1\left(\widehat{\mathcal{X}}^M_{\tau_{k-1}}, \mathcal{X}_{\tau_{k-1}}\right) \lesssim 2RN'\sqrt{d} \cdot \sqrt{\overline{\beta}^2 L_\varepsilon^2 dM\Delta t^2}.$$

Similarly, the DDPM bound (Lemma C.3) also does not contain the score approximation error:

$$\mathrm{KL}\left(f_{\theta^*}(\cdot, T)_\sharp \mathcal{X}_T, \mathcal{X}_\varepsilon\right) \lesssim \overline{\beta}^2 L_\varepsilon^2 Td\Delta t,$$

which leads to

$$W_1\left(f_{\theta^*}(\cdot, T)_\sharp \mathcal{X}_T, \mathcal{X}_\varepsilon\right) \lesssim \sqrt{d} \cdot \sqrt{\overline{\beta}^2 L_\varepsilon^2 Td\Delta t}$$

according to Lemma B.4. In the next step, we need to bound the Lipschitz constant L_ε since Assumption 4.1 is no longer applicable here.

Lemma D.2. *For the mixture of Gaussian distribution $\frac{1}{n}\sum_{j=1}^n \mathcal{N}(\mathbf{x}^j, \sigma^2\boldsymbol{I})$, we denote $\widehat{p}$ as its density. Assume $\|\mathbf{x}^j\|_2 \leqslant R_0$ for $\forall j \in [n]$, then its score function $\nabla \log \widehat{p}(\cdot)$ is L-Lipschitz continuous. Here $L = \max(R_0^2/\sigma^4, 1/\sigma^2)$. Furthermore, it leads to $L_\varepsilon = 4R_0^2/(\varepsilon^2\underline{\beta}^2)$.*

Proof. For the score function of $\frac{1}{n}\sum_{j=1}^{n}\mathcal{N}(\mathbf{x}^j,\sigma^2)$, it has the following formulation:

$$s(\mathbf{x}) = \frac{\sum_{j=1}^{n} -\frac{\mathbf{x}-\mathbf{x}^j}{\sigma^2}\exp\left(-\frac{\|\mathbf{x}-\mathbf{x}^j\|^2}{2\sigma^2}\right)}{\sum_{j=1}^{n}\exp\left(-\frac{\|\mathbf{x}-\mathbf{x}^j\|^2}{2\sigma^2}\right)} = -\sum_{j=1}^{n}\frac{\mathbf{x}-\mathbf{x}^j}{\sigma^2}\cdot p^j = -\frac{1}{\sigma^2}\mathbf{x} + \frac{1}{\sigma^2}\sum_{j=1}^{n}p^j\mathbf{x}^j.$$

Here,

$$p^j = \frac{\exp\left(-\frac{\|\mathbf{x}-\mathbf{x}^j\|^2}{2\sigma^2}\right)}{\sum_{k=1}^{n}\exp\left(-\frac{\|\mathbf{x}-\mathbf{x}^k\|^2}{2\sigma^2}\right)} \quad \forall j \in [n].$$

The Jacobian matrix

$$\begin{aligned}
\frac{\mathrm{d}s(\mathbf{x})}{\mathrm{d}\mathbf{x}} &= -\frac{1}{\sigma^2}\mathrm{Id} + \frac{1}{\sigma^2}\sum_{j=1}^{n}\mathbf{x}^j\cdot\left(\frac{\mathrm{d}p^j}{\mathrm{d}\mathbf{x}}\right)^\top = -\frac{1}{\sigma^2}\mathrm{Id} - \frac{1}{\sigma^2}\sum_{j=1}^{n}\mathbf{x}^j\cdot\left(p^j\cdot\frac{\mathbf{x}-\mathbf{x}^j}{\sigma^2} + p^j s(\mathbf{x})\right)^\top \\
&= -\frac{1}{\sigma^2}\mathrm{Id} + \frac{1}{\sigma^4}\sum_{j=1}^{n}p^j\mathbf{x}^j\mathbf{x}^{j\top} - \frac{1}{\sigma^2}\left(\sum_{j=1}^{n}p^j\mathbf{x}^j\right)\cdot\left(s(\mathbf{x}) + \frac{1}{\sigma^2}x\right)^\top \qquad (32) \\
&= -\frac{1}{\sigma^2}\mathrm{Id} + \frac{1}{\sigma^4}\sum_{j=1}^{n}p^j\mathbf{x}^j\mathbf{x}^{j\top} - \frac{1}{\sigma^4}\left(\sum_{j=1}^{n}p^j\mathbf{x}^j\right)\cdot\left(\sum_{j=1}^{n}p^j\mathbf{x}^j\right)^\top.
\end{aligned}$$

Therefore, we have:

$$-\frac{1}{\sigma^2}\mathrm{Id} \preceq \frac{\mathrm{d}s(\mathbf{x})}{\mathrm{d}\mathbf{x}} \preceq \frac{1}{\sigma^4}\sum_{j=1}^{n}p^j\mathbf{x}^j\mathbf{x}^{j\top}.$$

Notice that when $\|\mathbf{x}^j\|_2 \leqslant R_0$ for all $j \in [n]$:

$$\left\|\sum_{j=1}^{n}p^j\mathbf{x}^j\mathbf{x}^{j\top}\right\|_2 \leqslant \sum_{j=1}^{n}p^j\left\|\mathbf{x}^j\mathbf{x}^{j\top}\right\|_2 = \sum_{j=1}^{n}p^j\|\mathbf{x}^j\|_2^2 \leqslant R_0^2.$$

Finally, we can conclude that $s(\mathbf{x})$ is L-Lipschitz continuous for $L = \max(1/\sigma^2, R_0^2/\sigma^4)$. Furthermore, for L_ε, the Lipschitz continuity of $\nabla\log\widehat{p}_t(\cdot)$ for $t\in[\varepsilon,T]$, since

$$\sigma(\varepsilon)^2 = 1 - m(\varepsilon)^2 \geqslant 1 - \exp(\underline{\beta}\varepsilon) > \underline{\beta}\varepsilon/2,$$

we have $L_\varepsilon = 4R_0^2/(\underline{\beta}^2\varepsilon^2)$. □

After combining these conclusions together, we notice that:

$$\begin{aligned}
W_1\left(f_{\widehat{\theta}}(\cdot,T)_\sharp\mathcal{N}(0,\boldsymbol{I}), p_{\mathrm{data}}\right) &\leqslant W_1\left(f_{\widehat{\theta}}(\cdot,T)_\sharp\mathcal{N}(0,\boldsymbol{I}), f_{\widehat{\theta}}(\cdot,T)_\sharp\mathcal{X}_T\right) + W_1\left(f_{\widehat{\theta}}(\cdot,T)_\sharp\mathcal{X}_T, f_{\theta^*}(\cdot,T)_\sharp\mathcal{X}_T\right) \\
&\quad + W_1\left(f_{\theta^*}(\cdot,T)_\sharp\mathcal{X}_T, \mathcal{X}_\varepsilon\right) + W_1(\mathcal{X}_\varepsilon, p_{\mathrm{data}}) \\
&\lesssim R\cdot W_1(\mathcal{N}(0,\boldsymbol{I}),\mathcal{X}_T) + 2RN'\sqrt{d}\cdot\sqrt{\overline{\beta}^2L_\varepsilon^2dM\Delta t^2} \\
&\quad + \sqrt{d}\cdot\sqrt{\overline{\beta}^2L_\varepsilon^2Td\Delta t} + W_1(\mathcal{X}_\varepsilon, X_\varepsilon) + W_1(X_\varepsilon, p_{\mathrm{data}}).
\end{aligned}$$

Finally, we apply Lemma C.2, 5.5, C.4, and have:

$$\begin{gathered}
W_1(X_\varepsilon, p_{\mathrm{data}}) \lesssim \sqrt{d\overline{\beta}\varepsilon},\ \ \mathbb{E}\left[W_1(\mathcal{X}_\varepsilon, X_\varepsilon)\right] \leqslant \mathbb{E}\left[W_1(\widehat{p_{\mathrm{data}}}, p_{\mathrm{data}})\right] \lesssim n^{-1/d} \\
W_1(\mathcal{N}(0,\boldsymbol{I}),\mathcal{X}_T) \leqslant W_1(\mathcal{N}(0,\boldsymbol{I}),X_T) + W_1(X_T,\mathcal{X}_T) \lesssim \sqrt{d}\exp(-\underline{\beta}T/2) + n^{-1/d}.
\end{gathered}$$

To sum up, it holds that:

$$\begin{aligned}
\mathbb{E}\left[W_1\left(f_{\widehat{\theta}}(\cdot,T)_\sharp\mathcal{N}(0,\boldsymbol{I}), p_{\mathrm{data}}\right)\right] &\lesssim \sqrt{d}R\exp\left(-\underline{\beta}T/2\right) + R\cdot n^{-1/d} + d\overline{\beta}L_\varepsilon\cdot\frac{T}{\sqrt{M}} + \sqrt{d\overline{\beta}\varepsilon} \\
&\lesssim \sqrt{d}R\exp\left(-\underline{\beta}T/2\right) + R\cdot n^{-1/d} + \frac{d\overline{\beta}R_0^2}{\underline{\beta}^2\varepsilon^2}\cdot\frac{T}{\sqrt{M}} + \sqrt{d\overline{\beta}\varepsilon}, \qquad (33)
\end{aligned}$$

which comes to our conclusion of the main theorem 4.2.