On the Generalization of Diffusion Model

Mingyang Yi[1], **Jiacheng Sun**[1], **Zhenguo Li**[1]
[1]Huawei Noah's Ark Lab
{yimingyang2,sunjiacheng1,li.zhenguo}@huawei.com

Abstract

The diffusion probabilistic generative models are widely used to generate high-quality data. Though they can synthetic data that does not exist in the training set, the rationale behind such generalization is still unexplored. In this paper, we formally define the generalization of the generative model, which is measured by the mutual information between the generated data and the training set. The definition originates from the intuition that the model which generates data with less correlation to the training set exhibits better generalization ability. Meanwhile, we show that for the empirical optimal diffusion model, the data generated by a deterministic sampler are all highly related to the training set, thus poor generalization. This result contradicts the observation of the trained diffusion model's (approximating empirical optima) extrapolation ability (generating unseen data). To understand this contradiction, we empirically verify the difference between the sufficiently trained diffusion model and the empirical optima. We found, though obtained through sufficient training, there still exists a slight difference between them, which is critical to making the diffusion model generalizable. Moreover, we propose another training objective whose empirical optimal solution has no potential generalization problem. We empirically show that the proposed training objective returns a similar model to the original one, which further verifies the generalization ability of the trained diffusion model.

1 Introduction

The technique of generative model is capable of synthetic data from the target distribution, which has been well-developed in recent years e.g., VAE (Kingma and Welling, 2013), GAN (Goodfellow et al., 2014), and denoise diffusion probabilistic model (DDPM) (Song et al., 2020; Ho et al., 2020) etc. Among all these methods, the diffusion model has recently attracted great attention due to its capability of generating high-quality data that does not exist in the training set. However, some recent works (Somepalli et al., 2022; Carlini et al., 2023) have empirically shown that the diffusion model tends to generate data that is combined with the parts of data in the training set. This phenomenon threatens the application of the diffusion model in the aspect of privacy, as it may leak user's data (Carlini et al., 2023).

Ideally, the generative model should be capable of generating data from the underlying target distribution, but with less dependence on training data (so that extrapolating). Inspired by this intuition, we define the excess risk of the generative model which measures its performance of it. In contrast to the existing literature (Goodfellow et al., 2014; Arjovsky et al., 2017; Ho et al., 2020; Song et al., 2020), which only focuses on the quality of generated data, the defined excess risk also considers the generalization of the model. Concretely, our excess risk can be decomposed as the optimization error and the generalization error. The optimization error is explained as a distance between the distribution of generated data and the target one, which is the most commonly used metric to evaluate the generative model (Kingma and Welling, 2013). On the other hand, the generalization error cares about the "extrapolation" of the model, which intuitively is the correlation

between generated data and the training set. Owing to this, the generalization error is defined as the mutual information (Duchi, 2016) between them.

With the defined excess risk to measure the performance of the generative model, we apply it to check the quality of the diffusion model. As the model is trained by minimizing an empirical noise prediction problem (Song et al., 2020; Ho et al., 2020), we first analyze its empirical optimal solution. We show the solution can converge to the one with guaranteed optimization error. However, due to the formulation of the solution, generating data with deterministic update rule (Song et al., 2022; Lu et al., 2022) will generate data highly related to the training set, which results in poor generalization. Thus, as the sufficiently trained neural network can converge to the global minima of training objective (Allen-Zhu et al., 2019; Du et al., 2019), we are motivated to explore whether the poor generalization transfers to the well-trained diffusion model.

Fortunately, the empirical optimal solution has an explicit formulation, so we can directly compare it with the well-trained model. We empirically find that though the two models are close in each time step, the slight existing difference caused by optimization bias is critical for the diffusion model to generalize. This observation suggests that the neural network has the "regularization" property brought by the training stage (Zhang et al., 2021). We propose another training objective to verify the conclusion to get the diffusion model. The empirical optima of the proposed objective is shown to fix the generalization problem of the original one. We compare the models trained by the proposed and original objectives. The empirical results indicate that the two models have similar outputs, so we conclude that the potential generalization problem of diffusion can be obviated during the training of neural networks.

2 Related Work

Generalization of Generative Model. The classical generalization theory in prediction measures the gap between the model's performance on training and test data (Duchi, 2016; Vapnik, 1999; Yi et al., 2022). However, as the learned generative model does not take training data as input, the classical generalization theory does directly applied. To the best of our knowledge, (Arora et al., 2017) explore the generalization of GAN, while their definition measures the gap between population distance and empirical distance of the target and generated distributions. However, this notation is inconsistent with the intuition that a generalizable model can generate data that does not exist in the training set.

We measure generalization by correlating the generated and training data. The criterion is consistent with the intuition of generalization of the generative model, as we claimed in Section 1. The idea also originates from the informatic-generalization bound (Xu and Raginsky, 2017; Yi et al., 2023; Bu et al., 2020; Lopez and Jog, 2018), which says the correlation decides the generalization of the prediction problem between the model and training set.

Denoising Diffusion Probabilistic Model. The milestone work (Sohl-Dickstein et al., 2015) constructs a general formulation of the denoising diffusion probabilistic model, then specializes it by Gaussian and Binomial noises. By developing the Gaussian framework (diffusion model), (Song et al., 2020; Ho et al., 2020) obtain remarkable high-quality generated data. Thus, for the diffusion model, the left is verifying its generalization property. Though (Somepalli et al., 2022; Carlini et al., 2023) shows there are some generated samples that are quite similar to training data which may threaten the privacy of the diffusion model, our results show that the diffusion model can obviate memorizing training data (Somepalli et al., 2022; Carlini et al., 2023).

On the other hand, to get the diffusion model, we usually minimize the problems of noise prediction (Ho et al., 2020; Song et al., 2020) or data prediction (Cao et al., 2022; Gu et al., 2022). We propose to minimize the "previous points" to get a diffusion model, and we prove the proposed objective can obviate the potential generalization problem of the diffusion model.

3 Excess Risk of Generative Model

In this section, we formally define the excess risk (Yi et al., 2022) of the generative model, which evaluates the performance of it. Let training set $\boldsymbol{S} = \{\boldsymbol{x}_0^i\}_{i=1}^n$ be the n i.i.d. samples from target distribution P_0 with bounded support $\mathcal{X}$. The parameterized generative model $f_{\boldsymbol{\theta_S}}(\cdot)$ with $\boldsymbol{\theta_S}$ related

to the training set S transforms the variable v to the generated data $z = f_{\theta_S}(v)$ such that $z \sim Q_{\theta_S}$, where the v can be easily sampled e.g., Gaussian (Kingma and Welling, 2013; Goodfellow et al., 2014).

Intuitively, the ideal generative model is making Q_{θ_S} close to the target distribution P_0, but $z \sim Q_{\theta_S}$ is less related to training set S so that it generalize. The latter obviates the model generates data via memorizing the training set. For example, taking Q_{θ_S} as empirical distribution will generate data only from the training set. Though such Q_{θ_S} can converge to target distribution (Wainwright, 2019a), it clearly can not generalize. The following is the former definition of excess risk.

Definition 1 (Excess Risk)**.** *Let $z^j \sim Q_{\theta_S}$ generated by model f_{θ_S}, then the excess risk of f_{θ_S} is*

$$d_{\mathcal{F}}(Q_{\theta_S}, P_0) = \sup_{g \in \mathcal{F}} \left| \mathbb{E}_S \left[\limsup_{m \to \infty} \frac{1}{m} \sum_{j=1}^{m} g(z^j, S) - \mathbb{E}_{x \sim P_0}[g(x, S)] \right] \right|, \tag{1}$$

where $\mathcal{F} = \{g(x, S) : g(x, S) \in C(\mathcal{X}, \mathcal{X}^n)\}$.

Our definition originates from the probabilistic distance named integral probability metric (IPM) which is defined as

$$d_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} |\mathbb{E}_P[X] - \mathbb{E}_Q[X]|. \tag{2}$$

Clearly, only if Q_{θ_S} is close to P_0 for any $g(x, S) \in \mathcal{F}$, we can infinitely sample z and taking average to approximate $\mathbb{E}_{P_0}[g(x, S)]$. The correlation between z and S is induced by making $g(x, S)$ take S as input so that the correlation between z and S is involved in the excess risk. For example, the ideal model is making z^j independent with S, if $Q_{\theta_S} = P_0$, then $\limsup_{m \to \infty} \frac{1}{m} \sum_{j=1}^{m} g(z^j, S) \to \mathbb{E}_{z \sim P_0}[g(z, S)]$, and the excess risk becomes zero. The following theorem which is proved in Appendix A formulates the excess risk as an IPM.

Theorem 1. *If the generated data z^j in* (1) *are conditional independent with each other, given the training set S, and $\mathcal{F}$ has countable dense set under L_∞ distance, then the excess risk* (1) *becomes*

$$d_{\mathcal{F}}(Q_{\theta_S}, P_0) = \sup_{g \in \mathcal{F}} \left| \mathbb{E}_S \left[\mathbb{E}_{z \sim Q_{\theta_S}}[g(z, S)] - \mathbb{E}_{x \sim P_0}[g(x, S)] \right] \right|. \tag{3}$$

The conditional independence can be satisfied by many of generative models, e.g., GAN, diffusion model, VAE. Thus we explore the excess risk under such conditions in the sequel. At first glance, we can decompose it as

$$
\begin{aligned}
d_{\mathcal{F}}(Q_{\theta_S}, P_0) &\leq \sup_{g \in \mathcal{F}} \left| \mathbb{E}_{S, S'} \left[\mathbb{E}_{z \sim Q_{\theta_S}}[g(z, S)] - \mathbb{E}_{z \sim Q_{\theta_S}}[g(z, S')] \right] \right| \\
&\quad + \sup_{g \in \mathcal{F}} \left| \mathbb{E}_{S, S'} \left[\mathbb{E}_{z \sim Q_{\theta_S}}[g(z, S')] - \mathbb{E}_{z \sim P_0}[g(z, S)] \right] \right| \\
&\leq \underbrace{D_{\mathcal{F}}(P_{z_{\theta_S} \times S}, P_{z_{\theta_S}} \times P_S)}_{\text{generalization error}} + \underbrace{D_{\mathcal{F}}(Q_{\theta_S}, P_0)}_{\text{optimization error}},
\end{aligned}
\tag{4}
$$

where $D_{\mathcal{F}}(P, Q)$ is IPM defined in (2), and S' is another data set from P_0 independent with S. We explain the two terms in the above inequality. At first glance, the optimization error measures the distance between of generated distribution and the target one, which is the classical metric to evaluate the quality of generated data, e.g., JS-divergence (Goodfellow et al., 2014), KL-divergence (Kingma and Welling, 2013; Song et al., 2021), and Wasserstein distance (Arjovsky et al., 2017). On the other hand, the generalization error term measures the distance between union distribution $P_{z_{\theta_S} \times S}$ and $P_{z_{\theta_S}} \times P_S$. This is decided by the correlation between z_{θ_S} and training set S, which intuitively represents the generalization ability of the generative model. A similar correlation has been well explored in informatic-generalization theory (Xu and Raginsky, 2017; Rodríguez Gálvez et al., 2021). In their works, the generalization error of the prediction problem is decided by probabilistic distance with z substituted by the learned parameters. Finally, we make several examples to illustrate our excess risk in Appendix B.

As the generalization error should be influenced by the number of samples (Vapnik, 1999). To reduce such influence, we have the following proposition, in which we also link the generalization term to practical mutual information whose definition can be found in (Duchi, 2016).

Proposition 1. *Suppose $g(z, S) \in \mathcal{F}$ takes the form of $\frac{1}{n} \sum_{i=1}^{n} f(z, x_0^i)$ such that $\mathbb{E}_{Q_{\theta_S} \times P_0}[\exp f(z, x)] < \infty$ and $|f(z, x)| \leq M$, then*

$$d_{\mathcal{F}}(Q_{\theta_S}, P_0) \leq \sqrt{\frac{M^2}{n} I(z_{\theta_S}, S)} + d_{\mathcal{F}_{P_0}}(Q_{\theta_S}, P_0), \tag{5}$$

where $\mathcal{F}_{P_0} = \{\mathbb{E}_{\boldsymbol{x} \sim P_0}[f(\boldsymbol{z}, \boldsymbol{x})] : |f(\boldsymbol{z}, \boldsymbol{x})| \leq M; \mathbb{E}_{Q_{\boldsymbol{\theta}_{\boldsymbol{S}}} \times P_0}[\exp f(\boldsymbol{z}, \boldsymbol{x})] < \infty\}$ and $d_{\mathcal{F}_{P_0}}(Q_{\boldsymbol{\theta}_{\boldsymbol{S}}}, P_0) \leq \max\{D_{KL}(P_0, Q_{\boldsymbol{\theta}_{\boldsymbol{S}}}), D_{KL}(Q_{\boldsymbol{\theta}_{\boldsymbol{S}}}, P_0)\}$.

The proof of this theorem is in Appendix A. As can be seen, when we restrict the estimated term $g(\boldsymbol{z}, \boldsymbol{S})$ as the average over the training set, the generalization can be related to the number of training samples, which is consistent with our common sense. Besides that, the generalization error is decided by the mutual information between generated data and the training set.

4 Excess Risk of Diffusion Model

As we defined the excess risk to evaluate the generative model in Section 3, we apply it to the diffusion model in the sequel.

4.1 Revisiting Diffusion Model

As in (Ho et al., 2020), take $\boldsymbol{x}_0 \sim P_0$, and construct a forward process $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T\}$ such that $\boldsymbol{x}_{t+1} \mid \boldsymbol{x}_t \sim \mathcal{N}(\sqrt{1 - \beta_t}\boldsymbol{x}_t, \beta_t \boldsymbol{I})$, with $\beta_t > 0$ is variance schedule. By simple computation, we get

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}_t, \tag{6}$$

where $\boldsymbol{\epsilon}_t$ is a standard Gaussian noise independent with $\boldsymbol{x}_0$ and $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{1 \leq s \leq t} \alpha_s$. As can be seen, by properly designing β_t, the forward process obtains $\boldsymbol{x}_T$ that is close to a standard Gaussian distribution. Then to reversely generate $\boldsymbol{x}_0$, we can consider a reversed Markov process $\boldsymbol{x}_t$ such that $Q_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t), \Sigma_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t))$. Since $\boldsymbol{x}_T \approx \mathcal{N}(0, \boldsymbol{I})$, we can get $\boldsymbol{x}_{t-1}$ by iteratively sampling from $Q_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$, starting with a $\boldsymbol{x}_T$ sampled from standard Gaussian. To get transition probability $Q_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$, consider the constructed variational bound of maximal likelihood loss

$$\mathbb{E}_{\boldsymbol{x}_0 \sim P_0}\left[-\log Q_{\boldsymbol{\theta}}(\boldsymbol{x}_0)\right] \leq \mathbb{E}_P\left[-\log \frac{Q_{\boldsymbol{\theta}}(\boldsymbol{x}_{0:T})}{P(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)}\right]$$

$$= C + \mathbb{E}\left[\sum_{t>1} \underbrace{D_{KL}(P(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \parallel Q_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t))}_{L_{t-1}} \underbrace{-\log Q_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)}_{L_0}\right], \tag{7}$$

where C is a constant independent with $\boldsymbol{\theta}$. The update rule of $Q_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ can be obtained via minimizing $L_{\text{vb}} = \sum_{t=0}^{T-1} L_t$. By Bayes's rule, we have $P(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\beta}_t \boldsymbol{I})$ with

$$\tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t}\boldsymbol{x}_0 + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}\boldsymbol{x}_t; \qquad \tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}\beta_t, \tag{8}$$

Then we can explicitly get the optimal solution for each of L_{t-1} by selecting proper $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ and $\Sigma_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$. We have the following proposition proved in Appendix C to characterize the transition probability kernel $Q_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ for $t > 1$. On the other hand, as in (Song et al., 2022), the transition probability kernel of $Q_{\boldsymbol{\theta}}(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)$ is usually set as the mean in (9).

Proposition 2. For $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ with enough functional capacity, then

$$\underset{\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)}{\arg\min} L_{t-1} = \tilde{\boldsymbol{\mu}}_t\left(\boldsymbol{x}_t, \mathbb{E}\left[\boldsymbol{x}_0 \mid \boldsymbol{x}_t\right]\right); \qquad \underset{\Sigma_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)}{\arg\min} L_{t-1} = \tilde{\beta}_t. \tag{9}$$

In the widely used denoising diffusion probabilistic model (DDPM (Ho et al., 2020)) the transition rule is

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}^*(\boldsymbol{x}_t, t)\right); \qquad \Sigma_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) = \tilde{\beta}_t, \tag{10}$$

where $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}^*(\boldsymbol{x}_t, t)$ is a parameterized model such that $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}^* = \inf_{\boldsymbol{\epsilon}_{\boldsymbol{\theta}}} \mathbb{E}_{\boldsymbol{x}_t, \boldsymbol{\epsilon}_t}[\|\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \boldsymbol{\epsilon}_t\|^2]$. According to the optimality of conditional expectation under minimizing expected square loss (Banerjee et al., 2005), we know that the ideal

$$\boldsymbol{\epsilon}_{\boldsymbol{\theta}}^*(\boldsymbol{x}_t, t) = \mathbb{E}[\boldsymbol{\epsilon}_t \mid \boldsymbol{x}_t] = \mathbb{E}\left[\frac{1}{\sqrt{1 - \bar{\alpha}_t}}\boldsymbol{x}_t - \frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1 - \bar{\alpha}_t}}\boldsymbol{x}_0 \mid \boldsymbol{x}_t\right]. \tag{11}$$

By plugging this into (10), we get $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ is exactly the proposed optimal $\tilde{\boldsymbol{\mu}}_t\left(\boldsymbol{x}_t, \mathbb{E}\left[\boldsymbol{x}_0 \mid \boldsymbol{x}_t\right]\right)$. Thus, the rationale of standard DDPM is matching $P(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0)$ by substituting $\boldsymbol{x}_0$ with conditional expectation $\mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t]$. On the other hand, Proposition 2 indicates that such substitution is optimal in terms of minimizing variational bound L_{vb}.

4

4.2 Excess Risk of Diffusion Model

We have pointed out the optimal transition rule of the diffusion model above. Next, we verify the excess risk of the diffusion model under such a rule to generate data. In practice, to approximate the model $\epsilon_{\boldsymbol{\theta}}^*$ after (10), we minimize the following empirical counterpart of noise prediction problem $\inf_{\epsilon_{\boldsymbol{\theta}}} \mathbb{E}_{\boldsymbol{x}_t, \epsilon_t} \left[\|\epsilon_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \epsilon_t\|^2 \right]$.

$$\inf_{\epsilon_{\boldsymbol{\theta}}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\epsilon_t} \left[\left\| \epsilon_{\boldsymbol{\theta}}(\sqrt{\bar{\alpha}_t} \boldsymbol{x}_0^i + \sqrt{1 - \bar{\alpha}_t} \epsilon_t, t) - \epsilon_t \right\|^2 \right]. \tag{12}$$

The following two theorems explore the excess risk of the optima of (12).

Theorem 2. *Suppose the model $\epsilon_{\boldsymbol{\theta}}(\cdot, \cdot)$ has enough functional capacity, let $\epsilon_{\boldsymbol{\theta}_S}^*(\boldsymbol{x}, t)$ be any optima of* (12)*, then*

$$\epsilon_{\boldsymbol{\theta}_S}^*(\boldsymbol{x}, t) = \frac{\boldsymbol{x}}{\sqrt{1 - \bar{\alpha}_t}} - \left(\frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1 - \bar{\alpha}_t}} \right) \sum_{i=1}^{n} \frac{\exp \left(-\frac{\|\boldsymbol{x} - \sqrt{\bar{\alpha}_t} \boldsymbol{x}_0^i\|^2}{2(1 - \bar{\alpha}_t)} \right) \boldsymbol{x}_0^i}{\sum_{i=1}^{n} \exp \left(-\frac{\|\boldsymbol{x} - \sqrt{\bar{\alpha}_t} \boldsymbol{x}_0^i\|^2}{2(1 - \bar{\alpha}_t)} \right)}. \tag{13}$$

Then if the transition rule of DDPM satisfies (10) *as in (Ho et al., 2020), we have*

$$I(\boldsymbol{x}_0, \boldsymbol{S}) \leq \frac{(1 - \beta_1) R^2}{2 \beta_1^2} + \sum_{t=2}^{T} \frac{\bar{\alpha}_t R^2}{2(1 - \bar{\alpha}_{t-1})^2} \tag{14}$$

where $\boldsymbol{x}_0$ is generated by the model, then the generalization error in Proposition 1 is upper bounded.

The proof of this theorem is in Appendix C.1. The theorem indicates that the empirical optima of the noise prediction problem can have guaranteed generalization error when the R.H.S. of the above inequality is small. This happens when $1/\beta_1$ is not extremely large, which requires when constructing noisy data $\{\boldsymbol{x}_t\}$ the first $\boldsymbol{x}_1$ should be pretty noisy according to (18).

Next, we use the following theorem to indicate that such empirical optima also converge to the optimal model $\mathbb{E}[\epsilon_t \mid \boldsymbol{x}_t]$ as discussed in (11). Thus its ability to generate high-quality data is also guaranteed, as $\mathbb{E}[\epsilon_t \mid \boldsymbol{x}_t]$ minimizes L_{vb}, which is an upper bound of KL-divergence between generated distribution and target one (measures optimization error). The following theorem is proved in C.2.
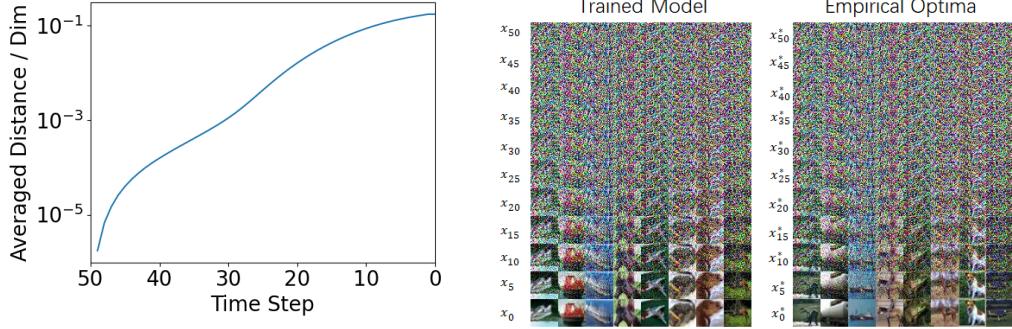
Theorem 3. *Let $\epsilon_{\boldsymbol{\theta}_S}^*(\cdot, \cdot)$ be the model defined in* (13)*, then for any t, and $\boldsymbol{x}_t$ with bounded norm, we have $\epsilon_{\boldsymbol{\theta}_S}^*(\boldsymbol{x}_t, t) \xrightarrow{P} \mathbb{E}[\epsilon_t \mid \boldsymbol{x}_t]$.*

Combining Theorem 2 and 3, we conclude that for a training set with a sufficiently large number, the DDPM can have guaranteed excess risk (under small β_1) so that generating high-quality data with a small dependence on the training set. However, the transition rule of DDPM is low efficient owing to a large T in practice, e.g., 1000 in (Ho et al., 2020). Because getting every $\boldsymbol{x}_t$ during generation requires taking a forward propagation of learned model $\epsilon_{\boldsymbol{\theta}_S}$, which takes plenty of computational costs. Researchers have proposed a deterministic reverse process (e.g., DDIM (Song et al., 2022)), which can generate high-quality data with fewer steps during its reverse process.

Unfortunately, as can be seen in (13), the empirical optima $\epsilon_{\boldsymbol{\theta}_S}^*(\boldsymbol{x}, t)$ takes the form of a linear combination of difference between the $\boldsymbol{x}$ and training set. Thus, any deterministic reverse process to generate $\boldsymbol{x}_t$ will make the generated data highly dependent on the training set, then poor generalization. The formal results is stated in the following proposition.

Proposition 3. *If the transition rule of the diffusion model takes the form of $\boldsymbol{x}_{t-1} = f(\epsilon_{\boldsymbol{\theta}_S}^*, \boldsymbol{x}_t, t)$ for some deterministic f. Then the generalization error of the diffusion model is infinity.*

To clarify the poor generalization, we take DDIM (Song et al., 2022) as an example. The $\boldsymbol{x}_{t-1}$ in DDIM is generating via a linear combination of $\boldsymbol{x}_t$ and $\epsilon_{\boldsymbol{\theta}_S}^*(\boldsymbol{x}_t, t)$, which results in the generated $\boldsymbol{x}_0$ must be a linear combination of training set. Clearly, we do not want such generated data as they only depend on the training set. Compared with DDPM, the guaranteed generalization of DDPM (10) originates the injected noise during generating process, which decreases the dependence between $\boldsymbol{x}_t$ and the training set.

5

(a) Averaged distance $\|\boldsymbol{x}_t - \boldsymbol{x}_t^*\|^2$ per dimension.

(b) Generated $\boldsymbol{x}_t$ and $\boldsymbol{x}_t^*$

Figure 1: The first figure is the averaged distance $\|\boldsymbol{x}_t - \boldsymbol{x}_t^*\|$ per dimension ($3{\times}32{\times}32$) over 50k samples of generated `CIFAR10`. The second figure randomly samples a batch of $\boldsymbol{x}_t$ and $\boldsymbol{x}_t^*$ with the same $\boldsymbol{x}_T = \boldsymbol{x}_T^*$ and $T = 50$.

The rationale for causing such a problem is the optimal model $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}^*$ in (11) involves conditional expectation $\mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t]$, and we require training set to estimate it. Since

$$\mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t] = \int_{\mathcal{X}} \boldsymbol{x}_0 P(\boldsymbol{x}_0 \mid \boldsymbol{x}_t) d\boldsymbol{x}_0 = \int_{\mathcal{X}} \boldsymbol{x}_0 \frac{P(\boldsymbol{x}_t \mid \boldsymbol{x}_0)}{P(\boldsymbol{x}_t)} P(\boldsymbol{x}_0) d\boldsymbol{x}_0 \approx \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_0^i \frac{P(\boldsymbol{x}_t \mid \boldsymbol{x}_0^i)}{\hat{P}(\boldsymbol{x}_t)}, \quad (15)$$

with $\hat{P}(\boldsymbol{x}_t)$ as a proper estimation to $P(\boldsymbol{x}_t)$, the estimator can be easily highly related to the training set. This can be verified by combining (11) and (13).

5 The Optimization Bias Improves Generalization

As we have claimed above, the empirical optima have a potential generalization problem. Unfortunately, this problem may be transferred to the sufficiently trained model as it approximates the empirical optima. Thus in this section, we explore whether the sufficiently trained model has generalization problem.

5.1 The Optimization Bias Regularizes Diffusion Model

Fortunately, we have the explicit formulation of empirical optima as in (13). Thus we can directly compare it with a sufficiently trained model. As we have claimed, when generating data with a deterministic reverse process, the empirical optima may generate data highly related training set. We explore it and verify whether this happens to the trained model.
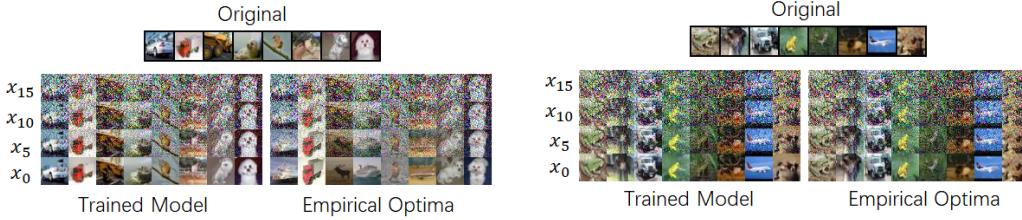
Following the pipeline in (Ho et al., 2020), we train a deep neural network, i.e., Unet (Ronneberger et al., 2015) on an image data set `CIFAR10` (Krizhevsky and Hinton, 2009) to verify the difference. We use 50-steps deterministic reverse process DDIM [1] as in (Song et al., 2022) such that

$$\boldsymbol{x}_{t+1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{\boldsymbol{x}_t - \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}_{\boldsymbol{\theta}(\boldsymbol{x}_t, t)}}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1}} \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) \quad (16)$$

to generate data. Let $\boldsymbol{x}_t$ and $\boldsymbol{x}_t^*$ respectively be the data generated by our trained model and the empirical optima. That means substituting $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ in the above equation with the trained model and empirical optima. We randomly sample 50K standard Gaussian and feed them into our trained and empirical optimal diffusion models. To check the difference, we summarize the averaged l_2-distance $\|\boldsymbol{x}_t - \boldsymbol{x}_t^*\|^2$ per dimension over the 50K iterates in Figure 1a. We also randomly sample some iterates $\{\boldsymbol{x}_t\}$ and $\{\boldsymbol{x}_t^*\}$ to visualize the difference in Figure 1b.

As can be seen, the distance between $\boldsymbol{x}_t$ and $\boldsymbol{x}_t^*$ increased with time step t. This is a natural result, as there is a gap between the trained model and the empirical optima owing to the bias brought by the

---

[1]The 50-step DDIM used here follows the one in (Song et al., 2022). After training a diffusion model build on $\boldsymbol{x}_{0:1000}$, then using $\boldsymbol{x}_{\lceil ci \rceil}$ with $i = 1, \cdot, 50$ and $c = 25$, as the new $\{\boldsymbol{x}_t\}$.

(a) The generted data starting from noisy data in test set of `CIFAR10`.

(b) Generted data started from noisy data in the training set.

Figure 2: The data in the two top figures are x_0 respectively from test and training sets of `CIFAR10`. The bottom are the data generated by the trained model (left) and empirical optima (right).

optimization process. Then the difference will cumulatively increase, resulting in different generated data as shown in Figure 1b. Thus we can conclude that the optimization bias regularizes the trained model to perfectly fit empirical optima, which instead potentially obviates the generalization problem.

5.2 The Optimization Bias Helps Extrapolating

Though the optimization process implicitly regularizes the trained model to generate data with the one generated by empirical optima. We should examine whether the two models will generate data that existed in the training set. Unfortunately, we observe that nearly all data generated by empirical optima exist in the training set, which also verifies our conclusion that the model has a generalization problem. On the other hand, for the trained model, we also compare the nearest data in the training set with its generated data to examine its generalization. Fortunately, we found that nearly all data does not appear in the training set. Thus, the optimization bias guarantees the extrapolation ability of the model. This phenomenon is shown in Appendix E.1.

To further verify the extrapolation ability of the trained diffusion model, we explicitly show that it can generate data that does not exist in the training set. Instead of starting the reverse process from x_{50}, we use x_{15} ($x_{15} = 0.6678x_0 + 0.7743\epsilon_t$) as an initial point so that we can check the generating process more clear. The x_0 we choose to x_{15} are from the test set, and a batch of generated data is in Figure 2b. As can be seen, the trained diffusion model nearly recovers the original data from the noisy ones. Nevertheless, for empirical optima, though starting from the same points, it can not recover the original unseen test data. By the way, decreasing the total sampling steps does change the result since the accumulated bias depends on the distance between x_T and x_0 (see (18)) which does influence by the number of sampling steps.

It has been observed in (Carlini et al., 2023) that the trained diffusion model occasionally generates data close to the one in the training set. Even though we think this does not threaten the extrapolation ability of the diffusion model. We conduct another experiment to explain such a phenomenon. Similar to the generating process in Figure 2b, we generated data by the diffusion model and empirical optima, starting from x_{15} but with x_0 drawn from the training set.

A batch of generated data is in Figure 2a. As can be seen, both the trained diffusion model and empirical optima can recover the original data from the noisy one. This explains why the diffusion model generates data in the training set. The generating happens if the reverse process moves to x_t around noisy data close to the one potentially constructed by the training set (like the x_{15}) when t is close to zero. The repeating generation is because the gap between generated data caused by the optimization bias of the trained diffusion model does not accumulate enough to regularize the training process. However, we think such repeating could hardly happen as x_t locates in high-dimensional space, so noisy data generated by the training set is sparse in its support (Wainwright, 2019b). Oppositely, with enough accumulated bias, we observe that such a phenomenon does not happen when taking $t = 50$ as in Figure 1b even though with x_{50} generated by the training set. We verify it in Appendix E.1.

Finally, we point out that this empirically observed phenomenon also holds for reverse process DDPM (Ho et al., 2020). As we have shown in Theorem 2, the generalization problem is resolved when β_1 is large, which does not hold for the one of DDPM ($\beta_1 = 0.0001$).

6 Estimating Previous Status Improves Generalization

As we have pointed out in (15), the potentially broken generalization property of the diffusion model originates from estimating $\mathbb{E}[\epsilon_t \mid \boldsymbol{x}_t]$ (equivalent to estimating $\mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t]$), which may lead the generated data highly related to the training set. Though this phenomenon can be mitigated by the optimization bias. We propose another training objective to get a diffusion model and generate data. Unlike the one of (12), the empirical optima of our proposed training objective mitigate the potential generalization problem.

Actually, we can rewrite the Proposition 2 such that

Proposition 4. *For $\boldsymbol{\mu_\theta}(\boldsymbol{x}_t, t)$ with enough functional capacity, then*

$$\underset{\boldsymbol{\mu_\theta}(\boldsymbol{x}_t,t)}{\arg\min} L_{t-1} = \mathbb{E}[\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t]; \qquad \underset{\Sigma_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)}{\arg\min} L_{t-1} = \tilde{\beta}_t. \tag{17}$$

As can be seen, in contrast to the transition probability rule in (9), the new rule does not involve $\mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t]$, so that it potentially obviates the generalization problem. Naturally, we may consider solving $\inf_{\boldsymbol{x_\theta}} \mathbb{E}[\|\boldsymbol{x}_{t-1} - \boldsymbol{x_\theta}(\boldsymbol{x}_t, t)\|^2]$ ($\boldsymbol{x_\theta}(\cdot, \cdot)$ is the parameterized diffusion model) to get $\mathbb{E}[\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t]$, as it is the solution of the minimization problem. However, practically, we found that the minimizing $\mathbb{E}[\|\boldsymbol{x}_{t-1} - \boldsymbol{x_\theta}(\boldsymbol{x}_t, t)\|^2]$ is unstable. We speculate this is due to the $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t-1}$ are so close which makes $\boldsymbol{x_\theta}(\boldsymbol{x}_t, t)$ rapidly converges to identity map of $\boldsymbol{x}_t$ for each t.

Owing to the aforementioned training problem, we consider another method to estimate the $\mathbb{E}[\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t]$. Suppose that

$$\boldsymbol{x}_t = \sqrt{\frac{\bar{\alpha}_t}{\bar{\alpha}_s}}\boldsymbol{x}_s + \sqrt{1 - \frac{\bar{\alpha}_t}{\bar{\alpha}_s}}\boldsymbol{\xi}_{t,s}, \tag{18}$$

and $\bar{\alpha}_t/\bar{\alpha}_s = r_{t,s}, \boldsymbol{\xi}_{t,s} \sim \mathcal{N}(0, \boldsymbol{I})$. Then

$$\mathbb{E}[\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t] = \frac{1}{\sqrt{r_{t,t-1}}}\boldsymbol{x}_t - \sqrt{\frac{1 - r_{t,t-1}}{r_{t,t-1}}}\mathbb{E}[\boldsymbol{\xi}_{t,t-1} \mid \boldsymbol{x}_t]. \tag{19}$$

Thus estimating $\mathbb{E}[\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t]$ is equivalent to estimating $\mathbb{E}[\boldsymbol{\xi}_{t,t-1} \mid \boldsymbol{x}_t]$. To get it, we have the following lemma, which is known as Tweedie's formula (Efron, 2011).

Lemma 1. *For and $s < t$, we have* $\frac{\mathbb{E}[\boldsymbol{\xi}_{t,s}|\boldsymbol{x}_t]}{\sqrt{1-r_{t,s}}} = \frac{\mathbb{E}[\boldsymbol{\xi}_{t,t-1}|\boldsymbol{x}_t]}{\sqrt{1-r_{t,t-1}}} = -\nabla_{\boldsymbol{x}_t} \log P_t(\boldsymbol{x}_t).$

From the above lemma, we know that estimating $\mathbb{E}[\boldsymbol{\xi}_{t,t-1} \mid \boldsymbol{x}_t]$ is equivalent to estimate $\mathbb{E}[\boldsymbol{\xi}_{t,s} \mid \boldsymbol{x}_t]$ for any $0 \geq s < t$, but the difference between $\boldsymbol{x}_t$ and $\boldsymbol{x}_s$ can be large when s is far away from t. We empirically find that a large gap benefits the optimization process. Thus our training objective becomes

$$\inf_{\boldsymbol{\xi_\theta}} \sum_{t=1}^{T} \mathbb{E}_s \left[\mathbb{E}_{\boldsymbol{x}_s,\boldsymbol{\xi}_{t,s}} \left[\left\| \frac{\boldsymbol{\xi}_{t,s}}{\sqrt{1 - r_{t,s}}} - \boldsymbol{\xi_\theta}(\sqrt{r_{t,s}}\boldsymbol{x}_s + \sqrt{1 - r_{t,s}}\boldsymbol{\xi}_{t,s}, t) \right\|^2 \right] \right], \tag{20}$$
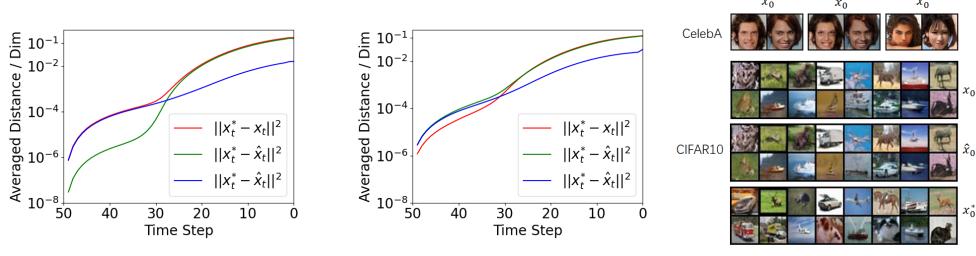
where s follows any distribution, e.g., uniform in $\{0, \cdots, T - 1\}$, and $\boldsymbol{\xi_\theta}$ is the final parameterized diffusion model. This can be done as for any specific t and s, the problem of minimizing $\mathbb{E}_{\boldsymbol{x}_s,\boldsymbol{\xi}_{t,s}} [\|\boldsymbol{\xi}_{t,s}/\sqrt{1 - r_{t,s}} - \boldsymbol{\xi_\theta}(\sqrt{r_{t,s}}\boldsymbol{x}_s + \sqrt{1 - r_{t,s}}\boldsymbol{\xi}_{t,s}, t)\|^2]$ has common global optima $\mathbb{E}[\boldsymbol{\xi}_{t,t-1} \mid \boldsymbol{x}_t]/\sqrt{1 - r_{t,t-1}}$ due to Lemma 1 and (Banerjee et al., 2005). Practically, let us consider the empirical counterpart of the above problem such that

$$\inf_{\boldsymbol{\xi_\theta}} \sum_{t=1}^{T} \mathbb{E}_s \left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\xi}_{t,s}} \left[\left\| \frac{\boldsymbol{\xi}_{t,s}}{\sqrt{1 - r_{t,s}}} - \boldsymbol{\xi_\theta}(\sqrt{r_{t,s}}\boldsymbol{x}_s^i + \sqrt{1 - r_{t,s}}\boldsymbol{\xi}_{t,s}, t) \right\|^2 \right] \right] \tag{21}$$

The $\{\boldsymbol{x}_s^i\}$ is generated through training set that follows the distribution of $\boldsymbol{x}_t$. The objective is actually equivalent to the (7) in reverse-SDE (Song et al., 2020) but substituting $\boldsymbol{x}_0$ with $\boldsymbol{x}_s$ as discussed in D. The following proposition gives the empirical optimal of (21).

Proposition 5. *Suppose the model $\boldsymbol{\xi_\theta}(\cdot, \cdot)$ has enough functional capacity the optimal solution of (21) is*

$$\boldsymbol{\xi}_{\boldsymbol{\theta}_S}^*(\boldsymbol{x}, t) = \sum_{i=1}^{n} \frac{\mathbb{E}_s \left[\left(\frac{1}{2\pi(1-r_{t,s})} \right)^{\frac{d}{2}} \exp\left(-\frac{\|\boldsymbol{x}-\sqrt{r_{t,s}}\boldsymbol{x}_s^i\|^2}{2(1-r_{t,s})} \right) \left(\frac{\boldsymbol{x}-\sqrt{r_{t,s}}\boldsymbol{x}_s^i}{1-r_{t,s}} \right) \right]}{\sum_{i=1}^{n} \mathbb{E}_s \left[\left(\frac{1}{2\pi(1-r_{t,s})} \right)^{\frac{d}{2}} \exp\left(-\frac{\|\boldsymbol{x}-\sqrt{r_{t,s}}\boldsymbol{x}_s^i\|^2}{2(1-r_{t,s})} \right) \right]}. \tag{22}$$

(a) Averaged distance per dimension on `CIFAR10`.

(b) Averaged distance per dimension on `CelebA`

(c) Comparasion of $\boldsymbol{x}_0, \hat{\boldsymbol{x}}_0, \boldsymbol{x}_0^*$.

Figure 3: The comparisons of $\boldsymbol{x}_t, \hat{\boldsymbol{x}}_t, \boldsymbol{x}_t^*$, where they are respectively generated by diffusion models trained by (12), (21), and the empirical optima (13).

As can be seen, in contrast to (13), the optimal solution $\boldsymbol{\xi}_{\boldsymbol{\theta}_S}^*(\boldsymbol{x}, t)$ does not highly relate to the training set. It involves $\{\boldsymbol{x}_s\}$ for series of s (depending on the distribution of s), and these $\{\boldsymbol{x}_s\}$ are noisy data generated by training set. Thus, despite the optimization bias discussed in Section 5, updating with (22) does not cause the potential generalization problem. The proof of this theorem is in Appendix D.

Similar to the Theorem 3, the proposed $\boldsymbol{\xi}_{\boldsymbol{\theta}_S}^*(\cdot, \cdot)$ also converges to its approximation target $\mathbb{E}[\boldsymbol{\xi}_{t-1} \mid \boldsymbol{x}_t]$, so that it has small optimization error when n is large enough. The result is illustrated in the following theorem, which is proved in D.

Theorem 4. *Let $\boldsymbol{\xi}_{\boldsymbol{\theta}_S}(\cdot, \cdot)$ be the model defined in* (22)*, then for any t and $\boldsymbol{x}_t$ with bounded norm, we have $\boldsymbol{\xi}_{\boldsymbol{\theta}_S}^*(\boldsymbol{x}_t, t) \xrightarrow{P} \mathbb{E}[\boldsymbol{\xi}_{t,t-1} \mid \boldsymbol{x}_t]/\sqrt{1 - r_{t,t-1}}.$*

7 Experiments

In Section 5, we empirically verify that although the empirical optimal diffusion model has a generalization problem, i.e., generating data from the training set. The optimization bias regularizes the trained diffusion model and enables it to generalize. In this section, we further verify the generalization capability of the diffusion model. We have shown in Section 6 proposed training objective (21) can obviate the potential generalization problem. Thus in this section, we empirically verify the difference between diffusion models trained by (12), (21), and the empirical optima (13).

Setup. Our experimental settings are similar to the ones in Section (5). That is, taking the reverse process as 50 steps DDIM (Song et al., 2022) to generate data. The diffusion models trained (12) and (21) are Unets (Ronneberger et al., 2015) with size depending on the dataset as in (Nichol and Dhariwal, 2021). To get s when training diffusion under our objective (21), for $k = t - s$, we first uniformly sample k from $1, \cdots, T-1$, then uniformly sample a s from $0, \cdots, T-k$. In addition, the sampled $\boldsymbol{x}_s$ during the training stage is generated by the training set according to (18). The other experimental settings follow the ones in (Ho et al., 2020).

Datasets. Our experiments are conducted on image datasets `CIFAR10` (Krizhevsky and Hinton, 2009), `CelebA` (Liu et al., 2015), which are all benchmark datasets with size 32×32 and 64×64.

Main Results. Similar to Section 5. Let $\boldsymbol{x}_t^*$, $\boldsymbol{x}_t$ and $\tilde{\boldsymbol{x}}_t$ respectively be the iterates generated by empirical optima, diffusion models trained by (12) and (21). For each model, we generate 50k series of iterates to compare the average difference (per dimension) between them. The comparisons are summarized in Figure 3. As can be seen, compared with $\boldsymbol{x}_t^*$, the iterates generated by trained diffusion models are pretty similar. This illustrates that though they are trained by different objectives, but the optimization bias pushes them towards a similar model with generalization ability. As the model trained by (21) does not have a potential generalization problem, the similarity between $\boldsymbol{x}_t$ and $\hat{\boldsymbol{x}}_t$ indicates the generalization ability of diffusion model trained by (21).

Some samples generated by the three models are in Figure 3c. As can be seen, the $\boldsymbol{x}_0$ and $\hat{\boldsymbol{x}}_0$ are visually close to each other, while $\hat{\boldsymbol{x}}_0$ are noisy compared with $\boldsymbol{x}_0$. In fact, $\hat{\boldsymbol{x}}_0$ exhibits higher FID

score (lower is better)(Heusel et al., 2017) than x_0 (evaluated as in (Ho et al., 2020)), that is 11.30 v.s. 3.17 on CIFAR10 and 108.73 v.s. 8.91 on CelebA. We speculate this is because $\hat{x}_t$ are more noisy, which improves generalization but increases the optimization error. This illustrates that there is a trade-off between the two errors. Thus when training the diffusion model, we should consider balancing them.

8 Conclusion

In this paper, we first formally define the excess risk of the generative model to evaluate it. The excess risk can be decomposed into optimization and generalization errors, which relate to the quality of generated data and the model's exploration ability, respectively. We mainly focus on exploring the generalization of the diffusion model. We verify that though the empirical optimal diffusion model has poor generalization, the optimization bias brought by the training stage of the diffusion model enables it to generate high quality, but meanwhile preserving the generalization ability.

References

Allen-Zhu, Z., Li, Y., and Song, Z. (2019). A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*.

Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. (2017). Generalization and equilibrium in generative adversarial nets (gans). In *International Conference on Machine Learning*.

Banerjee, A., Guo, X., and Wang, H. (2005). On the optimality of conditional expectation as a bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669.

Bu, Y., Zou, S., and Veeravalli, V. V. (2020). Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 1(1):121–130.

Cao, H., Tan, C., Gao, Z., Chen, G., Heng, P.-A., and Li, S. Z. (2022). A survey on generative diffusion model. Technical Report arXiv:2209.02646.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., and Wallace, E. (2023). Extracting training data from diffusion models. Preprint arXiv:2301.13188.

Du, S., Lee, J. D., Li, H., Wang, L., and Zhai, X. (2019). Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*.

Duchi, J. (2016). Lecture notes for statistics 311/electrical engineering 377. *URL: https://stanford. edu/class/stats311/Lectures/full_notes. pdf. Last visited on*, 2:23.

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

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*.

Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., and Guo, B. (2022). Vector quantized diffusion model for text-to-image synthesis. In *Conference on Computer Vision and Pattern Recognition*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational {Bayes}. In *International Conference on Learning Representations*.

Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *International Conference on Computer Vision*.

Lopez, A. T. and Jog, V. (2018). Generalization error bounds using wasserstein distances. In *2018 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE.

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

Nichol, A. Q. and Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*.

Rodríguez Gálvez, B., Bassi, G., Thobaben, R., and Skoglund, M. (2021). Tighter expected generalization error bounds via wasserstein distance.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer.

Shiryaev, A. N. (2016). *Probability-1*, volume 95. Springer.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. (2022). Diffusion art or digital forgery? investigating data replication in diffusion models. Preprint arXiv:2212.03860.

Song, J., Meng, C., and Ermon, S. (2022). Denoising diffusion implicit models. In *International Conference on Learning Representations*.

Song, Y., Durkan, C., Murray, I., and Ermon, S. (2021). Maximum likelihood training of score-based diffusion models.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2020). Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.

Van Der Vaart, A. W., Wellner, J. A., van der Vaart, A. W., and Wellner, J. A. (1996). *Weak convergence*. Springer.

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer science & business media.

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

Wainwright, M. J. (2019a). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

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

Xu, A. and Raginsky, M. (2017). Information-theoretic analysis of generalization capability of learning algorithms.

Yi, M., Wang, R., and Ma, Z.-M. (2022). Characterization of excess risk for locally strongly convex population risk. *Advances in Neural Information Processing Systems*.

Yi, M., Wang, R., Sun, J., Li, Z., and Ma, Z.-M. (2023). Breaking correlation shift via conditional invariant regularizer. In *The International Conference on Learning Representations*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115.

A Proofs in Section 3

Theorem 1. *If the generated data z^j in (1) are conditional independent with each other, given the training set S, and $\mathcal{F}$ has countable dense set under L_∞ distance, then the excess risk (1) becomes*

$$d_{\mathcal{F}}(Q_{\boldsymbol{\theta}_S}, P_0) = \sup_{g \in \mathcal{F}} \left| \mathbb{E}_S \left[\mathbb{E}_{z \sim Q_{\boldsymbol{\theta}_S}}[g(z, S)] - \mathbb{E}_{x \sim P_0}[g(x, S)] \right] \right|. \tag{3}$$

Proof. As given the training set S, the generated data z^j are conditional independent with each other. Then for any g, and a realization of training set S_0, we have that

$$\limsup_{m \to \infty} \frac{1}{m} g(z^j, S_0) = \mathbb{E}_{z \sim Q_{\boldsymbol{\theta}_{S_0}}}[g(z, S_0)], \qquad a.s., \tag{23}$$

where $a.s.$ means almost surely. Due to $\mathcal{X}$ is bounded, it has countable dense sets. Then for any dense set $\mathcal{X}_0^n$ of $\mathcal{X}^n$, we have the above equation holds for any $S_0 \in \mathcal{X}_0^n$ almost surely. Then for any $S \in \mathcal{X}^n$, due to the continuity of g w.r.t. S, and $\mathcal{X}_0^n$ is a dense subset of $\mathcal{X}^n$, we have

$$\begin{aligned}
\limsup_{m \to \infty} \frac{1}{m} g(z^j, S) - \mathbb{E}_{z \sim Q_{\boldsymbol{\theta}_S}}[g(z, S)] &= \limsup_{m \to \infty} \frac{1}{m} g(z^j, S) - \limsup_{m \to \infty} \frac{1}{m} g(z^j, S^\epsilon) \\
&\quad + \mathbb{E}_{z \sim Q_{\boldsymbol{\theta}_{S^\epsilon}}}[g(z, S^\epsilon)] - \mathbb{E}_{z \sim Q_{\boldsymbol{\theta}_S}}[g(z, S)] \\
&\leq 2\mathcal{O}(\epsilon), \qquad a.s.
\end{aligned} \tag{24}$$

where $S^\epsilon \in \mathcal{X}_0^n$ such that $\|S^\epsilon - S\| \leq \epsilon$. Then due to the arbitrary of ϵ, we get that

$$\mathbb{E}_S \left[\limsup_{m \to \infty} \frac{1}{m} g(z^j, S) \right] = \mathbb{E}_S \mathbb{E}_{z \sim Q_{\boldsymbol{\theta}_S}}[g(z, S)], \qquad a.s. \tag{25}$$

holds for any fixed g. For any countable dense set $\mathcal{F}_0$ of $\mathcal{F}$, we have

$$d_{\mathcal{F}_0}(Q_{\boldsymbol{\theta}_S}, P_0) = \sup_{g \in \mathcal{F}_0} \left| \mathbb{E}_S \left[\mathbb{E}_{z \sim Q_{\boldsymbol{\theta}_S}}[g(z, S)] - \mathbb{E}_{x \sim P_0}[g(x, S)] \right] \right|. \qquad a.s. \tag{26}$$

Then for any $\epsilon > 0$, there must exists dense set $\mathcal{F}_0$ such that

$$\begin{aligned}
&\left| d_{\mathcal{F}}(Q_{\boldsymbol{\theta}_S}, P_0) - \sup_{g \in \mathcal{F}} \left| \mathbb{E}_S \left[\mathbb{E}_{z \sim Q_{\boldsymbol{\theta}_S}}[g(z, S)] - \mathbb{E}_{x \sim P_0}[g(x, S)] \right] \right| \right| \\
&\leq |d_{\mathcal{F}}(Q_{\boldsymbol{\theta}_S}, P_0) - d_{\mathcal{F}_0}(Q_{\boldsymbol{\theta}_S}, P_0)| \\
&+ \left| d_{\mathcal{F}}(Q_{\boldsymbol{\theta}_S}, P_0) - \sup_{g \in \mathcal{F}} \left| \mathbb{E}_S \left[\mathbb{E}_{z \sim Q_{\boldsymbol{\theta}_S}}[g(z, S)] - \mathbb{E}_{x \sim P_0}[g(x, S)] \right] \right| \right| \\
&+ \left| \sup_{g \in \mathcal{F}} \left| \mathbb{E}_S \left[\mathbb{E}_{z \sim Q_{\boldsymbol{\theta}_S}}[g(z, S)] - \mathbb{E}_{x \sim P_0}[g(x, S)] \right] \right| - \sup_{g \in \mathcal{F}_0} \left| \mathbb{E}_S \left[\mathbb{E}_{z \sim Q_{\boldsymbol{\theta}_S}}[g(z, S)] - \mathbb{E}_{x \sim P_0}[g(x, S)] \right] \right| \right| \\
&\leq 3\epsilon.
\end{aligned} \tag{27}$$

Let us define event

$$A = \left\{ d_{\mathcal{F}}(Q_{\boldsymbol{\theta}_S}, P_0) \neq \sup_{g \in \mathcal{F}} \left| \mathbb{E}_S \left[\mathbb{E}_{z \sim Q_{\boldsymbol{\theta}_S}}[g(z, S)] - \mathbb{E}_{x \sim P_0}[g(x, S)] \right] \right| \right\}. \tag{28}$$

Then $A = \bigcup_{\epsilon > 0} A^\epsilon = \bigcup_{j=1}^\infty A^{\frac{1}{j}}$, with

$$A^\epsilon = \left\{ \left| d_{\mathcal{F}}(Q_{\boldsymbol{\theta}_S}, P_0) - \sup_{g \in \mathcal{F}} \left| \mathbb{E}_S \left[\mathbb{E}_{z \sim Q_{\boldsymbol{\theta}_S}}[g(z, S)] - \mathbb{E}_{x \sim P_0}[g(x, S)] \right] \right| \right| \geq \epsilon \right\}. \tag{29}$$

Due to the denseness of $\mathcal{F}_0$, we can get

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{\epsilon > 0} A^\epsilon \right) = \mathbb{P}\left(\bigcup_{j=1}^\infty A^{\frac{1}{j}} \right) \leq \sum_{j=1}^\infty \mathbb{P}(A^{\frac{1}{j}}) = 0, \tag{30}$$

where the last equality is due to (27). Thus we prove our result. $\qquad \square$

Proposition 1. *Suppose $g(z, S) \in \mathcal{F}$ takes the form of $\frac{1}{n}\sum_{i=1}^n f(z, x_0^i)$ such that $\mathbb{E}_{Q_{\boldsymbol{\theta}_S} \times P_0}[\exp f(z, x)] < \infty$ and $|f(z, x)| \leq M$, then*

$$d_{\mathcal{F}}(Q_{\boldsymbol{\theta}_S}, P_0) \leq \sqrt{\frac{M^2}{n} I(z_{\boldsymbol{\theta}_S}, S)} + d_{\mathcal{F}_{P_0}}(Q_{\boldsymbol{\theta}_S}, P_0), \tag{5}$$

where $\mathcal{F}_{P_0} = \{\mathbb{E}_{x \sim P_0}[f(z, x)] : |f(z, x)| \leq M; \mathbb{E}_{Q_{\boldsymbol{\theta}_S} \times P_0}[\exp f(z, x)] < \infty\}$ and $d_{\mathcal{F}_{P_0}}(Q_{\boldsymbol{\theta}_S}, P_0) \leq \max\{D_{KL}(P_0, Q_{\boldsymbol{\theta}_S}), D_{KL}(Q_{\boldsymbol{\theta}_S}, P_0)\}$.

Proof. Let us check the generalization error first. Due to the formulation of $g(\boldsymbol{z}, \boldsymbol{S})$, for any $\lambda > 0$

$$\lambda \sup_{g \in \mathcal{F}} \mathbb{E}_{\boldsymbol{S}, \boldsymbol{S}'} \left[\mathbb{E}_{\boldsymbol{z} \sim Q_{\boldsymbol{\theta}_{\boldsymbol{S}}}}[g(\boldsymbol{z}, \boldsymbol{S})] - \mathbb{E}_{\boldsymbol{z} \sim Q_{\boldsymbol{\theta}_{\boldsymbol{S}}}}[g(\boldsymbol{z}, \boldsymbol{S}')] \right] = \lambda \mathbb{E}_{\boldsymbol{S}, \boldsymbol{z} \sim Q_{\boldsymbol{\theta}_{\boldsymbol{S}}}} \left[\frac{1}{n} \sum_{i=1}^{n} \left(f(\boldsymbol{z}, \boldsymbol{x}_0^i) - \mathbb{E}_{\boldsymbol{S}'}[f(\boldsymbol{z}, \boldsymbol{x}_0^{i'})] \right) \right]$$

$$\leq D_{KL}(P_{\boldsymbol{z}_{\boldsymbol{\theta}_{\boldsymbol{S}}} \times \boldsymbol{S}}, P_{\boldsymbol{z}_{\boldsymbol{\theta}_{\boldsymbol{S}}}} \times P_{\boldsymbol{S}}) + \log \mathbb{E}_{\boldsymbol{S}', \boldsymbol{z} \sim Q_{\boldsymbol{\theta}_{\boldsymbol{S}}}} \left[\exp \left(\frac{\lambda}{n} \sum_{i=1}^{n} \left(f(\boldsymbol{z}, \boldsymbol{x}_0^i) - \mathbb{E}_{\boldsymbol{S}'}[f(\boldsymbol{z}, \boldsymbol{x}_0^{i'})] \right) \right) \right]$$

$$\stackrel{a}{\leq} I(\boldsymbol{z}_{\boldsymbol{\theta}_{\boldsymbol{S}}}, \boldsymbol{S}) + \frac{\lambda^2 M^2}{2n},$$

(31)

where inequality a is from the sub-Gaussian property (Duchi, 2016). By taking infimum of λ, and similarly applying the result to $\lambda < 0$, we prove an upper bound to the generalization error that is

$$D_{\mathcal{F}}(P_{\boldsymbol{z}_{\boldsymbol{\theta}_{\boldsymbol{S}}} \times \boldsymbol{S}}, P_{\boldsymbol{z}_{\boldsymbol{\theta}_{\boldsymbol{S}}}} \times P_{\boldsymbol{S}}) \leq \sqrt{\frac{M^2 I(\boldsymbol{z}_{\boldsymbol{\theta}_{\boldsymbol{S}}}, \boldsymbol{S})}{n}}.$$

(32)

On the other hand, for the optimization error, then

$$\sup_{g \in \mathcal{F}} \mathbb{E}_{\boldsymbol{S}, \boldsymbol{S}'} \left| \left[\mathbb{E}_{\boldsymbol{z} \sim Q_{\boldsymbol{\theta}_{\boldsymbol{S}}}}[g(\boldsymbol{z}, \boldsymbol{S}')] - \mathbb{E}_{\boldsymbol{z} \sim P_0}[g(\boldsymbol{z}, \boldsymbol{S})] \right] \right|$$

$$= \sup_{f} \left| \mathbb{E}_{\boldsymbol{x} \sim P_0} \left[\left(\mathbb{E}_{\boldsymbol{z} \sim Q_{\boldsymbol{\theta}_{\boldsymbol{S}}}}[f(\boldsymbol{z}, \boldsymbol{x})] - \mathbb{E}_{\boldsymbol{z} \sim P_0}[f(\boldsymbol{z}, \boldsymbol{x})] \right) \right] \right|$$

$$= d_{\mathcal{F}_{P_0}}(Q_{\boldsymbol{\theta}_{\boldsymbol{S}}}, P_0),$$

(33)

where $\mathcal{F}_{P_0} = \{ \mathbb{E}_{\boldsymbol{x} \sim P_0}[f(\boldsymbol{z}, \boldsymbol{x})] : |f(\boldsymbol{z}, \boldsymbol{x})| \leq M; \mathbb{E}_{Q_{\boldsymbol{\theta}_{\boldsymbol{S}}} \times P_0}[\exp f(\boldsymbol{z}, \boldsymbol{x})] < \infty \}$. Then due to the Donsker–Varadhan representation (Duchi, 2016), we have $d_{\mathcal{F}_{P_0}}(Q_{\boldsymbol{\theta}_{\boldsymbol{S}}}, P_0) \leq \max\{D_{KL}(P_0, Q_{\boldsymbol{\theta}_{\boldsymbol{S}}}), D_{KL}(Q_{\boldsymbol{\theta}_{\boldsymbol{S}}}, P_0)\}$, which proves our theorem. $\square$

B Some Examples of Excess Risk

To make our excess risk more practical, we use the following example to illustrate the effectiveness.

Example 1. *The excess risk of empirical distribution $Q_{\boldsymbol{\theta}_{\boldsymbol{S}}} = P_n$.*

According to Theorem 1, the excess risk of empirical distribution is

$$d_{\mathcal{F}}(P_n, P_0) = \left| \sup_{g \in \mathcal{F}} \mathbb{E}_{\boldsymbol{S}} \left[\frac{1}{n} \sum_{i=1}^{n} g(\boldsymbol{x}_0^i, \boldsymbol{S}) - \mathbb{E}_{\boldsymbol{x} \sim P_0}[g(\boldsymbol{x}, \boldsymbol{S})] \right] \right|$$

$$\geq \left| \mathbb{E}_{\boldsymbol{S}} \left[\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\|\boldsymbol{x}_0^i - \boldsymbol{x}^j\|^2} \right] - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{x} \sim P_0} \frac{1}{\|\boldsymbol{x} - \boldsymbol{x}_0^i\|^2} \right|$$

$$= \infty.$$

(34)

As can be seen, when we involve the generalization into excess risk, the empirical distribution has poor performance, which is consistent with our intuition. However, under the original metric, the empirical distribution can have great performance with sufficiently large n. This is because the existing evaluation (Kingma and Welling, 2013; Song et al., 2020; Arora et al., 2017) is usually probabilistic distance or divergence between generated and target distributions, e.g., Wasserstein distance or KL divergence. However, the classical results of empirical process (Van Der Vaart et al., 1996) indicate that the empirical distribution can converge P_0 under these metrics. This contradicts to our intuition that memorizing training data is a bad behavior for the generative model.

Next, let us present an example to exactly compute the excess risk of generative model.

Example 2. *Let i.i.d. training set $\boldsymbol{S}$ such that $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I})$. Our goal is using $\boldsymbol{S}$ to estimate $\boldsymbol{\mu}$ by $\hat{\boldsymbol{\mu}}$ and generate data $\boldsymbol{z} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}, \boldsymbol{I})$. Now let us check the generalization and optimization error of generated data $\boldsymbol{z}$.*

The classical way to get $\hat{\boldsymbol{\mu}}$ is minimizing the square loss $\frac{1}{n} \sum_{i=1}^{n} \|\hat{\boldsymbol{\mu}} - \boldsymbol{x}_i\|^2$, which is obtained by $\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i$. Thus, $\boldsymbol{z} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}, \boldsymbol{I})$. We consider the function class as $\mathcal{F} = \{f : |f| \leq M\}$ for some $M > 0$. As in the proof of proposition 1, the $D_{\mathcal{F}}(\cdot, \cdot)$ defined in (4) can be upper bounded by $\max\{D_{KL}(Q_{\hat{\boldsymbol{\mu}}}||P_0), D_{KL}(P_0||Q_{\hat{\boldsymbol{\mu}}})\}$ by applying Jensen's inequality and its definition. Thus, the

14

optimization error of $Q_{\hat{\boldsymbol{\mu}}}$ can be explicitly computed due to KL-divergence between two Gaussian distributions (Duchi, 2016) such that

$$D_{\mathcal{F}}(P_0, Q_{\hat{\boldsymbol{\mu}}}) \leq D_{KL}(Q_{\hat{\boldsymbol{\mu}}} \| P_0) = D_{KL}(P_0 \| Q_{\hat{\boldsymbol{\mu}}}) = \mathbb{E}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2] = \frac{d}{n}. \tag{35}$$

On the other hand, for the generalization error, similar to the proof of Proposition 1, for some standard Gaussian distribution $\boldsymbol{\xi}$, we have

$$
\begin{aligned}
D_{\mathcal{F}}(P_{\boldsymbol{z} \times \boldsymbol{S}}, Q_{\hat{\boldsymbol{\mu}}} \times P_{\boldsymbol{S}}) &\leq I(\boldsymbol{z}; \boldsymbol{S}) \\
&= \sum_{i=1}^n I(\boldsymbol{z}; \boldsymbol{x}_i \mid \boldsymbol{x}_{1:i-1}) \\
&= \sum_{i=1}^n I\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i + \boldsymbol{\xi}; \boldsymbol{x}_i \mid \boldsymbol{x}_{1:i-1}\right) \\
&= \sum_{i=1}^n \left(H\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i + \boldsymbol{\xi} \mid \boldsymbol{x}_{1:i-1}\right) - H\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i + \boldsymbol{\xi} \mid \boldsymbol{x}_{1:i}\right)\right) \\
&= H\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i + \boldsymbol{\xi}\right) + H(\boldsymbol{\xi}) \\
&= \frac{d}{2}\log\left(1 + \frac{1}{n}\right),
\end{aligned}
\tag{36}
$$

where the last equality is due to the entropy of Gaussian distribution (Duchi, 2016). Thus we respectively characterize the upper bounds of generalization and optimization errors.

C Proofs in Section 4

We prove a general result such that our Proposition 2 is a corollary of it. We first present the definition of exponential family distributions, which is adopted from (Duchi, 2016)

Definition 2 (Exponential Family Distributions). *The exponential family associated with the function $\phi(\cdot)$ is defined as the set of distributions with densities $Q_{\boldsymbol{\theta}}$, where*

$$Q_{\boldsymbol{\theta}}(\boldsymbol{x}) = \exp\left(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}) \rangle - A(\boldsymbol{\theta})\right), \tag{37}$$

and the function $A(\boldsymbol{\theta})$ is the log-partition-function defined as

$$A(\boldsymbol{\theta}) = \log \int_{\mathcal{X}} \exp\left(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}) \rangle\right) d\boldsymbol{x} \tag{38}$$

Before proving Proposition 2, we need the following lemma.

Lemma 2. *For densities functions $P(\cdot)$ and $Q_{\boldsymbol{\theta}}(\cdot)$, if $Q_{\boldsymbol{\theta}}(\cdot)$ is an exponential family variable, then*

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} D_{KL}(P \| Q_{\boldsymbol{\theta}}) = \nabla A^{-1}(\mathbb{E}_p[\phi(\boldsymbol{x})]), \tag{39}$$

and $\mathbb{E}_{Q_{\boldsymbol{\theta}^*}}[\phi(\boldsymbol{x})] = \mathbb{E}_P[\phi(\boldsymbol{x})]$, where $\nabla A^{-1}(\boldsymbol{\theta})$ is the inverse of $\nabla A(\boldsymbol{\theta})$, due to the convexity of $A(\boldsymbol{\theta})$

Proof. From the definition

$$
\begin{aligned}
D_{KL}(P \| Q_{\boldsymbol{\theta}}) &= \int_{\mathcal{X}} P(\boldsymbol{x}) \log P(\boldsymbol{x}) d\boldsymbol{x} - \int_{\mathcal{X}} P(\boldsymbol{x}) \log Q_{\boldsymbol{\theta}}(\boldsymbol{x}) d\boldsymbol{x} \\
&= \int_{\mathcal{X}} p(\boldsymbol{x}) \log P(\boldsymbol{x}) d\boldsymbol{x} - \int_{\mathcal{X}} P(\boldsymbol{x}) \left(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}) - A(\boldsymbol{\theta})\rangle\right) d\boldsymbol{x}.
\end{aligned}
\tag{40}
$$

Then minimizing $D_{KL}(P \| Q_{\boldsymbol{\theta}})$ is equivalent to maximizing $-\int_{\mathcal{X}} P(\boldsymbol{x})\left(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}) - A(\boldsymbol{\theta})\rangle\right) d\boldsymbol{x}$. According to Proposition 14.4 in (Duchi, 2016), $A(\boldsymbol{\theta})$ is a convex function w.r.t. $\boldsymbol{\theta}$, then let $Q_{\boldsymbol{\theta}^*}$ solves (39), we must have $\nabla A(\boldsymbol{\theta}^*) = \mathbb{E}_{\boldsymbol{x} \sim P}[\phi(\boldsymbol{x})]$. On the other hand, we have

$$\nabla A(\boldsymbol{\theta}) = \frac{\int_{\mathcal{X}} \exp\left(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}) \rangle\right) \phi(\boldsymbol{x}) d\boldsymbol{x}}{\int_{\mathcal{X}} \exp\left(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x}) \rangle\right) d\boldsymbol{x}} = \mathbb{E}_{\boldsymbol{x} \sim Q_{\boldsymbol{\theta}}}[\phi(\boldsymbol{x})], \tag{41}$$

which verifies the second conclusion. $\square$

Proposition 2. *For $\boldsymbol{\mu_\theta}(\boldsymbol{x}_t, t)$ with enough functional capacity, then*

$$\underset{\boldsymbol{\mu_\theta}(\boldsymbol{x}_t,t)}{\arg\min} L_{t-1} = \tilde{\boldsymbol{\mu}}_t\left(\boldsymbol{x}_t, \mathbb{E}\left[\boldsymbol{x}_0 \mid \boldsymbol{x}_t\right]\right); \qquad \underset{\Sigma_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)}{\arg\min} L_{t-1} = \tilde{\beta}_t. \tag{9}$$

Proof. The normal distribution is exponential family with the form $Q_{\boldsymbol{\theta},\Sigma}(\boldsymbol{x}) \propto \exp(\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle + 1/2\langle \boldsymbol{x}\boldsymbol{x}^\top, \Sigma \rangle)$, where Σ is the covariance matrix of $Q_{\boldsymbol{\theta},\Sigma}(\boldsymbol{x})$ and $\boldsymbol{\theta}$ is $\Sigma^{-1}\boldsymbol{\mu}$ with $\boldsymbol{\mu}$ is the mean of $Q_{\boldsymbol{\theta},\Sigma}(\boldsymbol{x})$. Then the result is a corollary of Lemma 2 due to the linearity of $\tilde{\boldsymbol{\mu}}(\boldsymbol{x}_0, \boldsymbol{x}_t)$ w.r.t. $\boldsymbol{x}_0$. $\square$

C.1 The Empirical Optima of Noise Prediction

Next we prove the Theorem 2.

Theorem 2. *Suppose the model $\boldsymbol{\epsilon_\theta}(\cdot, \cdot)$ has enough functional capacity, let $\boldsymbol{\epsilon}^*_{\boldsymbol{\theta}_S}(\boldsymbol{x}, t)$ be any optima of (12), then*

$$\boldsymbol{\epsilon}^*_{\boldsymbol{\theta}_S}(\boldsymbol{x}, t) = \frac{\boldsymbol{x}}{\sqrt{1-\bar{\alpha}_t}} - \left(\frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}\right) \sum_{i=1}^{n} \frac{\exp\left(-\frac{\|\boldsymbol{x}-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^i\|^2}{2(1-\bar{\alpha}_t)}\right)\boldsymbol{x}_0^i}{\sum_{i=1}^{n}\exp\left(-\frac{\|\boldsymbol{x}-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^i\|^2}{2(1-\bar{\alpha}_t)}\right)}. \tag{13}$$

Then if the transition rule of DDPM satisfies (10) as in (Ho et al., 2020), we have

$$I(\boldsymbol{x}_0, \boldsymbol{S}) \le \frac{(1-\beta_1)R^2}{2\beta_1^2} + \sum_{t=2}^{T} \frac{\bar{\alpha}_t R^2}{2(1-\bar{\alpha}_{t-1})^2} \tag{14}$$

where $\boldsymbol{x}_0$ is generated by the model, then the generalization error in Proposition 1 is upper bounded.

Proof. Let

$$\begin{aligned}
J(\boldsymbol{\epsilon_\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{\epsilon}_t}\left[\left\|\boldsymbol{\epsilon_\theta}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^i + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t, t) - \boldsymbol{\epsilon}_t\right\|^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}^d}\left\|\boldsymbol{\epsilon_\theta}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^i + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t, t) - \boldsymbol{\epsilon}_t\right\|^2\left(\frac{1}{2\pi}\right)^{\frac{d}{2}}\exp\left(-\frac{\|\boldsymbol{\epsilon}_t\|^2}{2}\right)d\boldsymbol{\epsilon}_t \\
&= \int_{\mathbb{R}^d}\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{\epsilon_\theta}(\boldsymbol{x}, t) - \frac{\boldsymbol{x}-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^i}{\sqrt{1-\bar{\alpha}_t}}\right\|^2\left(\frac{1}{2\pi}\right)^{\frac{d}{2}}\exp\left(-\frac{\|\boldsymbol{x}-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^i\|^2}{2(1-\bar{\alpha}_t)}\right)d\boldsymbol{x}
\end{aligned} \tag{42}$$

For any given $\boldsymbol{x}$, the optimization problem of minimizing $\boldsymbol{\epsilon_\theta}$ in the integral is a strongly convex problem w.r.t. $\boldsymbol{\epsilon_\theta}$. Thus it has single global minimum which can be obtained taking gradient to it such that

$$\begin{aligned}
0 &= \nabla_{\boldsymbol{\theta}}\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{\epsilon_\theta}(\boldsymbol{x}, t) - \frac{\boldsymbol{x}-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^i}{\sqrt{1-\bar{\alpha}_t}}\right\|^2\left(\frac{1}{2\pi}\right)^{\frac{d}{2}}\exp\left(-\frac{\|\boldsymbol{x}-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^i\|^2}{2(1-\bar{\alpha}_t)}\right) \\
&= \frac{2}{n}\sum_{i=1}^{n}\left(\boldsymbol{\epsilon_\theta}(\boldsymbol{x}, t) - \frac{\boldsymbol{x}-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^i}{\sqrt{1-\bar{\alpha}_t}}\right)\left(\frac{1}{2\pi}\right)^{\frac{d}{2}}\exp\left(-\frac{\|\boldsymbol{x}-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^i\|^2}{2(1-\bar{\alpha}_t)}\right),
\end{aligned} \tag{43}$$

which shows

$$\begin{aligned}
\boldsymbol{\epsilon}^*_{\boldsymbol{\theta}_S}(\boldsymbol{x}, t) &= \sum_{i=1}^{n}\frac{\exp\left(-\frac{\|\boldsymbol{x}-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^i\|^2}{2(1-\bar{\alpha}_t)}\right)}{\sum_{i=1}^{n}\exp\left(-\frac{\|\boldsymbol{x}-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^i\|^2}{2(1-\bar{\alpha}_t)}\right)}\left(\frac{\boldsymbol{x}-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^i}{\sqrt{1-\bar{\alpha}_t}}\right) \\
&= \frac{\boldsymbol{x}}{\sqrt{1-\bar{\alpha}_t}} - \left(\frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}\right)\sum_{i=1}^{n}\frac{\exp\left(-\frac{\|\boldsymbol{x}-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^i\|^2}{2(1-\bar{\alpha}_t)}\right)\boldsymbol{x}_0^i}{\sum_{i=1}^{n}\exp\left(-\frac{\|\boldsymbol{x}-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^i\|^2}{2(1-\bar{\alpha}_t)}\right)}
\end{aligned} \tag{44}$$

Next, we prove the claim of generalization, due to the Proposition 1, we should control the mutual information $I(\boldsymbol{x}_0, \boldsymbol{S})$, where $\boldsymbol{x}_0$ is obtained via

$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}^*_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\right) + \tilde{\beta}_t\boldsymbol{\xi}_t, \tag{45}$$

where $\boldsymbol{\xi}_t$ is a standard Gaussian that is independent of $\boldsymbol{x}_t$ and $\boldsymbol{S}$. Then by Data processing inequality (Xu and Raginsky, 2017),

$$I(\boldsymbol{x}_0; \boldsymbol{S}) \leq I(\boldsymbol{x}_{0:T}; \boldsymbol{S}) = I(\boldsymbol{x}_0; \boldsymbol{S} \mid \boldsymbol{x}_{1:T}) + I(\boldsymbol{x}_1; \boldsymbol{S} \mid \boldsymbol{x}_{2:T}) + \cdots + I(\boldsymbol{x}_T; \boldsymbol{S}). \tag{46}$$

Then for any $1 \leq t \leq T$,

$$I(\boldsymbol{x}_{t-1}; \boldsymbol{S} \mid \boldsymbol{x}_{t:T}) = H(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_{t:T}) - H(\boldsymbol{x}_{t-1} \mid \boldsymbol{S}, \boldsymbol{x}_{t:T}) = H(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) - H(\boldsymbol{x}_{t-1} \mid \boldsymbol{S}, \boldsymbol{x}_t), \tag{47}$$

where the last equality is due to the Markovian property of $\boldsymbol{x}_{t-1}$. Next we compute the two terms in the last equality. First, due to the definition of $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}^*$

$$\begin{aligned} H(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) &= H\left(\boldsymbol{x}_{t-1} - \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t \mid \boldsymbol{x}_t\right) \\ &= H\left(\tilde{\beta}_t\boldsymbol{\xi}_t + \left(\beta_t\frac{\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t}\right)\sum_{i=1}^{n} \frac{\exp\left(-\frac{\|\boldsymbol{x}-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^i\|^2}{2(1-\bar{\alpha}_t)}\right)\boldsymbol{x}_0^i}{\sum_{i=1}^{n}\exp\left(-\frac{\|\boldsymbol{x}-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^i\|^2}{2(1-\bar{\alpha}_t)}\right)} \mid \boldsymbol{x}_t\right). \end{aligned} \tag{48}$$

Then since $\boldsymbol{x}_t$ and $\boldsymbol{\xi}_t$ are independent we have

$$\begin{aligned} &\mathbb{E}\left[\left\|\tilde{\beta}_t\boldsymbol{\xi}_t + \left(\frac{\beta_t\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t}\right)\sum_{i=1}^{n} \frac{\exp\left(-\frac{\|\boldsymbol{x}-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^i\|^2}{2(1-\bar{\alpha}_t)}\right)\boldsymbol{x}_0^i}{\sum_{i=1}^{n}\exp\left(-\frac{\|\boldsymbol{x}-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^i\|^2}{2(1-\bar{\alpha}_t)}\right)}\right\|^2\right] \\ &= \tilde{\beta}_t^2 d + \frac{\beta_t^2\bar{\alpha}_t}{(1-\bar{\alpha}_t)^2}\mathbb{E}\left[\left\|\sum_{i=1}^{n} \frac{\exp\left(-\frac{\|\boldsymbol{x}-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^i\|^2}{2(1-\bar{\alpha}_t)}\right)\boldsymbol{x}_0^i}{\sum_{i=1}^{n}\exp\left(-\frac{\|\boldsymbol{x}-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^i\|^2}{2(1-\bar{\alpha}_t)}\right)}\right\|^2\right] \\ &\leq \tilde{\beta}_t^2 d + \frac{\beta_t^2\bar{\alpha}_t}{(1-\bar{\alpha}_t)^2}R^2, \end{aligned} \tag{49}$$

where R is the radius of data support $\mathcal{X}$. Due to Theorem 14.7 in (Duchi, 2016), that among all random variables X with $\mathbb{E}[\|X\|^2 \leq C]$, the Gaussian distribution $\mathcal{N}(0, \sqrt{C/d}I_d)$ has the largest entropy such that

$$H(\mathcal{N}(0, \sqrt{C/d}I_d)) = \frac{d}{2}\log\left(\frac{2\pi eC}{d}\right). \tag{50}$$

Combining this result with (49), we get

$$H(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) \leq \frac{d}{2}\log\left(2\pi e\left(\tilde{\beta}_t^2 + \frac{\beta_t^2\bar{\alpha}_t}{d(1-\bar{\alpha}_t)^2}R^2\right)\right). \tag{51}$$

On the other hand, due to the definition of $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}^*$,

$$H(\boldsymbol{x}_{t-1} \mid \boldsymbol{S}, \boldsymbol{x}_t) = H(\tilde{\beta}_t\boldsymbol{\xi}_t) = \frac{d}{2}\log(2\pi e\tilde{\beta}_t^2), \tag{52}$$

which implies

$$I(\boldsymbol{x}_{t-1}; \boldsymbol{S} \mid \boldsymbol{x}_{t:T}) \leq \frac{d}{2}\log\left(1 + \frac{\beta_t^2\bar{\alpha}_t}{d\tilde{\beta}_t^2(1-\bar{\alpha}_t)^2}R^2\right) \leq \frac{\beta_t^2\bar{\alpha}_t R^2}{2\tilde{\beta}_t^2(1-\bar{\alpha}_t)^2} = \frac{\bar{\alpha}_t R^2}{2(1-\bar{\alpha}_{t-1})^2}. \tag{53}$$

We should point out that when $t = 1$, the upper bounded in the above becomes $\frac{(1-\beta_1)R^2}{2\beta_1^2}$. Then we prove our result. $\qquad\square$

Remark 1. *As we have shown in main text, the ideal $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ is (11), and the empirical $\boldsymbol{\epsilon}_{\boldsymbol{\theta}_S}$ in (13) is approximating (11) as in (15). This conclusion can be easily verified due to $P(\boldsymbol{x}_t \mid \boldsymbol{x}_0) \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, \sqrt{1-\bar{\alpha}_t}\boldsymbol{I})$, by viewing the numerator of (13) as an unbiased estimator to $P(\boldsymbol{x}_t)$, which is $\hat{P}(\boldsymbol{x}_t)$ in (15).*

In the last of this subsection, we use the following proposition to show the generalization problem of empirical when the diffusion model takes deterministic update rule e.g., DDIM (Song et al., 2022), DPM-Solver (Lu et al., 2022).

Proposition 3. *If the transition rule of the diffusion model takes the form of $\boldsymbol{x}_{t-1} = f(\boldsymbol{\epsilon}^*_{\boldsymbol{\theta}_S}, \boldsymbol{x}_t, t)$ for some deterministic f. Then the generalization error of the diffusion model is infinity.*

Proof. As can be seen, the $\boldsymbol{\epsilon}^*_{\boldsymbol{\theta}_S}(\boldsymbol{x})$ is a linear combination of the difference between $\boldsymbol{x}$ and training set $\boldsymbol{S}$. Thus according to the transition rule $\boldsymbol{x}_{t-1} = f(\boldsymbol{\epsilon}^*_{\boldsymbol{\theta}_S}, \boldsymbol{x}_t, t)$, we know the generated data $\boldsymbol{x}_0$ only depends on $\boldsymbol{S}$ and $\boldsymbol{x}_T$. Due to the linear formulation of $\boldsymbol{\epsilon}^*_{\boldsymbol{\theta}_S}(\boldsymbol{x}, t)$, there exists $\boldsymbol{x}_0 = F(\boldsymbol{x}_T, \boldsymbol{S})$ with F does not degenerated w.r.t. $\boldsymbol{S}$. Thus

$$I(\boldsymbol{x}_0, \boldsymbol{S}) = I(F(\boldsymbol{x}_T, \boldsymbol{S}); \boldsymbol{S}) = I(F(\boldsymbol{x}_T, \boldsymbol{S}); \boldsymbol{S} \mid \boldsymbol{x}_T) + I(\boldsymbol{x}_T; \boldsymbol{S}) = I(F(\boldsymbol{x}_T, \boldsymbol{S}); \boldsymbol{S} \mid \boldsymbol{x}_T) = \infty, \quad (54)$$

which verifies our conclusion. □

The proposition indicates that though the deterministic update rule of diffusion model has improved sampling efficiency compared with the stochastic one (Song et al., 2022; Lu et al., 2022), but it potentially face the challenge of generalization.

C.2 Convergence of Empirical Minima

In this section, we prove the convergence result of empirical minima (13). Before proving Theorem 3, we give some notations and present some useful lemmas. Let us define

$$K_t(\boldsymbol{x}_t, \boldsymbol{x}_0) = \exp\left(-\frac{\|\boldsymbol{x} - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0\|^2}{2(1 - \bar{\alpha}_t)}\right);$$

$$f_{\boldsymbol{x}_0}(\boldsymbol{x}_t) = \left(\frac{1}{2\pi(1 - \bar{\alpha}_t)}\right)^{\frac{d}{2}} K_t(\boldsymbol{x}_t, \boldsymbol{x}_0); \quad (55)$$

$$f_{\boldsymbol{S}}(\boldsymbol{x}_t) = \frac{1}{n}\sum_{i=1}^n f_{\boldsymbol{x}_0^i}(\boldsymbol{x}_t).$$

Lemma 3. *The function $f_{\boldsymbol{S}}(\boldsymbol{x}_t)$ and $P_t(\boldsymbol{x}_t)$ is $\left(\frac{1}{2\pi(1-\bar{\alpha}_t)}\right)^{\frac{d}{2}} \sqrt{\frac{1}{e(1-\bar{\alpha}_t)}}$-Lipschitz continuous and their gradients are all $\left(\frac{1}{2\pi(1-\bar{\alpha}_t)}\right)^{\frac{d}{2}} \left(\frac{2+e}{e(1-\bar{\alpha}_t)}\right)$-Lipschitz continuous.*

Proof. Due to the definition of $K_t(\boldsymbol{x}_t, \boldsymbol{x}_0)$, we have

$$\|\nabla_{\boldsymbol{x}_t} K_t(\boldsymbol{x}_t, \boldsymbol{x}_0)\| = -\exp\left(-\frac{\|\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0\|^2}{2(1 - \bar{\alpha}_t)}\right)\left\|\frac{\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0}{1 - \bar{\alpha}_t}\right\| \leq \sqrt{\frac{1}{e(1 - \bar{\alpha}_t)}}, \quad (56)$$

where we use the inequality $axe^{-\frac{ax^2}{2}} \leq \sqrt{a/e}$, then the Lipschitz continuity of $f_{\boldsymbol{S}}(\boldsymbol{x}_t)$ and $P_t(\boldsymbol{x}_t)$ are directly obtained since $P_t(\boldsymbol{x}_t) = \int_{\mathbb{R}^d} \left(\frac{1}{2\pi(1-\bar{\alpha}_t)}\right)^{\frac{d}{2}} P_0(\boldsymbol{x}_0)d\boldsymbol{x}_0$. On the other hand,

$$\nabla^2_{\boldsymbol{x}_t\boldsymbol{x}_t} K_t(\boldsymbol{x}_t, \boldsymbol{x}_0) = \exp\left(-\frac{\|\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0\|^2}{2(1 - \bar{\alpha}_t)}\right)\left(\frac{\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0}{1 - \bar{\alpha}_t}\right)\left(\frac{\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0}{1 - \bar{\alpha}_t}\right)^{\top}$$
$$+ \left(\frac{1}{1 - \bar{\alpha}_t}\right)\exp\left(-\frac{\|\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0\|^2}{2(1 - \bar{\alpha}_t)}\right)\boldsymbol{I}. \quad (57)$$

Thus for any $\boldsymbol{\xi} \in \mathbb{R}^d$ with $\|\boldsymbol{\xi}\| = 1$, we have

$$\sup_{\boldsymbol{\xi}: \|\boldsymbol{\xi}\|=1} \boldsymbol{\xi}^{\top} \nabla^2_{\boldsymbol{x}_y\boldsymbol{x}_t} K_t(\boldsymbol{x}_t, \boldsymbol{x}_0)\boldsymbol{\xi} = \exp\left(-\frac{\|\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0\|^2}{2(1 - \bar{\alpha}_t)}\right)\left\|\frac{\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0}{1 - \bar{\alpha}_t}\right\|^2$$
$$+ \left(\frac{1}{1 - \bar{\alpha}_t}\right)\exp\left(-\frac{\|\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0\|^2}{2(1 - \bar{\alpha}_t)}\right) \quad (58)$$
$$\leq \frac{2}{e(1 - \bar{\alpha}_t)} + \frac{1}{1 - \bar{\alpha}_t},$$

where we use the inequality $axe^{-\frac{ax}{2}} \leq 2e^{-1}$, which proves our second conclusion. □

18

The following lemma is an important transformation of conditional expectation which is Tweedie's Formula (Efron, 2011).

Lemma 4. *Suppose that $y \mid x \sim \mathcal{N}(\alpha x, \beta I)$, then $\mathbb{E}_{x|y}[x] = \frac{1}{\alpha}(y + \beta \nabla_v \log P(v))$.*

Theorem 3. *Let $\epsilon_{\theta_S}^*(\cdot, \cdot)$ be the model defined in (13), then for any t, and x_t with bounded norm, we have $\epsilon_{\theta_S}^*(x_t, t) \xrightarrow{P} \mathbb{E}[\epsilon_t \mid x_t]$.*

Proof. Due to Lemma 4, we have

$$
\begin{aligned}
\mathbb{E}[\epsilon_t \mid x_t] &= \mathbb{E}\left[\frac{1}{\sqrt{1-\bar{\alpha}_t}}x_t - \frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}x_0 \mid x_t\right] \\
&= \frac{1}{\sqrt{1-\bar{\alpha}_t}}x_t - \frac{1}{\sqrt{1-\bar{\alpha}_t}}\left(x_t + (1-\bar{\alpha}_t)\nabla_x \log P_t(x_t)\right) \\
&= -\sqrt{1-\bar{\alpha}_t}\nabla_x \log P_t(x_t).
\end{aligned}
\tag{59}
$$

Thus our goal is proving $\epsilon_{\theta_S}(x_t, t) \xrightarrow{P} -\sqrt{1-\bar{\alpha}_t}\nabla_x \log P_t(x_t)$.

$$
\begin{aligned}
\nabla_x \log P_t(x_t) &= \nabla_x P_t(x_t)/P_t(x_t) \\
&= \frac{\nabla_x \int_{\mathbb{R}^d} P_{t|0}(x_t \mid x_0)P_0(x_0)dx_0}{\int_{\mathbb{R}^d} P_{t|0}(x_t \mid x_0)P_0(x_0)}dx_0 \\
&= \frac{\mathbb{E}_{x_0}[\nabla_x P_{t|0}(x_t \mid x_0)]}{\mathbb{E}_{x_0}[P_{t|0}(x_t \mid x_0)]}.
\end{aligned}
\tag{60}
$$

Rewriting the (13) as

$$
\begin{aligned}
\epsilon_{\theta_S}^*(x_t, t) &= -\sqrt{1-\bar{\alpha}_t}\frac{\frac{1}{n}\sum_{i=1}^n \left(\frac{1}{2\pi(1-\bar{\alpha}_t)}\right)^{\frac{d}{2}} K_t(x_t, x_0^i)\left(\frac{x_t - \sqrt{\bar{\alpha}_t}x_0^i}{1-\bar{\alpha}_t}\right)}{\frac{1}{n}\sum_{i=1}^n \left(\frac{1}{2\pi(1-\bar{\alpha}_t)}\right)^{\frac{d}{2}} K_t(x_t, x_0^i)} \\
&= -\sqrt{1-\bar{\alpha}_t}\frac{\nabla_x f_S(x_t)}{f_S(x_t)},
\end{aligned}
\tag{61}
$$

where $K_t(x_t, x_0^i) = \exp\left(-\frac{\|x - \sqrt{\bar{\alpha}_t}x_0^i\|^2}{2(1-\bar{\alpha}_t)}\right)$. Then, what left is to show that $\frac{\nabla_x f_S(x_t)}{f_S(x_t)} \xrightarrow{P} \nabla_x \log P_t(x_t)$.

As can be seen the numerator and denominator of the above equation are respectively empirical estimator of the numerator and denominator of the one in (60). Then, both of them are consistency so that we get the conclusion. To check this, we have

$$
\mathbb{E}_S[f_S(x_t)] = \mathbb{E}_{x_0}[P_{t|0}(x_t \mid x_0)] = P_t(x_t).
\tag{62}
$$

Note that for any $S^{i'}$ equals to S expected $x_0^{i'} \neq x_0^i$, then for any $D > 0$

$$
\begin{aligned}
\sup_{x_t : \|x_t\| < D}(f_S(x_t) - P_t(x_t)) - \sup_{x_t : \|x_t\| < D}\left(f_{S^{i'}}(x_t) - P_t(x_t)\right) &\leq \sup_{x_t : \|x_t\| < D}(f_S(x_t) - f_{S^{i'}}(x_t)) \\
&= \frac{1}{n}\left(\frac{1}{2\pi(1-\bar{\alpha}_t)}\right)^{\frac{d}{2}}\sup_{x_t}\left(K_t(x_t, x_0^i) - K_t(x_t, x_0^{i'})\right) \\
&\leq \frac{1}{n}\left(\frac{1}{2\pi(1-\bar{\alpha}_t)}\right)^{\frac{d}{2}}.
\end{aligned}
\tag{63}
$$

Thus by McDiarmid's inequality, we must have

$$
\mathbb{P}\left(\left|\sup_{x_t : \|x_t\| \leq D}(f_S(x_t) - P_t(x_t)) - \mathbb{E}\left[\sup_{x_t : \|x_t\| \leq D}(f_S(x_t) - P_t(x_t))\right]\right| \geq \epsilon\right) \leq \exp\left(-2N(2\pi(1-\bar{\alpha}_t))^d\epsilon^2\right).
\tag{64}
$$

Thus $\sup_{x_t : \|x_t\| \leq D}(f_S(x_t) - P_t(x_t)) \xrightarrow{P} \mathbb{E}\left[\sup_{x_t : \|x_t\| \leq D}(f_S(x_t) - P_t(x_t))\right]$. For any x_t, y_t with norm smaller than D and $\lambda > 0$, let

$$
D_j = \mathbb{E}\left[f_S(x_t) - P_t(x_t) \mid x_0^{1:j}\right] - \mathbb{E}\left[f_S(x_t) - P_t(x_t) \mid x_0^{1:j-1}\right].
\tag{65}
$$

Then $f_{\boldsymbol{S}}(\boldsymbol{x}_t) - P_t(\boldsymbol{x}_t) = \sum_{j=1}^{n} D_j$. Let

$$U_j = \sup_{\boldsymbol{x}_0^j} \mathbb{E}\left[f_{\boldsymbol{S}}(\boldsymbol{x}_t) - P_t(\boldsymbol{x}_t) \mid \boldsymbol{x}_0^{1:j}\right] - \mathbb{E}\left[f_{\boldsymbol{S}}(\boldsymbol{x}_t) - P_t(\boldsymbol{x}_t) \mid \boldsymbol{x}_0^{1:j-1}\right];$$

$$L_j = \inf_{\boldsymbol{x}_0^j} \mathbb{E}\left[f_{\boldsymbol{S}}(\boldsymbol{x}_t) - P_t(\boldsymbol{x}_t) \mid \boldsymbol{x}_0^{1:j}\right] - \mathbb{E}\left[f_{\boldsymbol{S}}(\boldsymbol{x}_t) - P_t(\boldsymbol{x}_t) \mid \boldsymbol{x}_0^{1:j-1}\right], \tag{66}$$

we have $L_j \le D_j \le U_j$. Thus

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{S}}\left[\exp\left(\lambda\left[f_{\boldsymbol{S}}(\boldsymbol{x}_t) - P_t(\boldsymbol{x}_t)\right]\right)\right] &= \mathbb{E}_{\boldsymbol{S}}\left[\exp\left(\lambda \sum_{j=1}^{n} D_j\right)\right] \\
&= \mathbb{E}_{\boldsymbol{S}}\left[\mathbb{E}\left[\exp\left(\lambda \sum_{j=1}^{n} D_j\right) \mid \boldsymbol{x}_0^{1:N-1}\right]\right] \\
&= \mathbb{E}_{\boldsymbol{S}}\left[\exp\left(\lambda \sum_{j=1}^{N-1} D_j\right) \mathbb{E}\left[\exp\left(\lambda D_N\right) \mid \boldsymbol{x}_0^{1:N-1}\right]\right] \\
&= \prod_{j=1}^{n} \mathbb{E}_{\boldsymbol{S}}\left[\mathbb{E}\left[\exp\left(\lambda D_j\right) \mid \boldsymbol{x}_0^{1:j-1}\right]\right] \\
&\le \exp\left(\sum_{j=1}^{n} \frac{\lambda^2 (U_j - L_j)^2}{8}\right),
\end{aligned}
\tag{67}
$$

where the last inequality is due to Azuma-Hoeffding's inequality (Duchi, 2016). On the other hand, we have

$$
\begin{aligned}
U_j - L_j &\le \sup_{\boldsymbol{x}_0^i, \boldsymbol{x}_0^{i'}} \frac{1}{n}\left[\left(f_{\boldsymbol{x}_0^i}(\boldsymbol{x}_t) - P_t(\boldsymbol{x}_t)\right) - \left(f_{\boldsymbol{x}_0^{i'}}(\boldsymbol{x}_t) - P_t(\boldsymbol{x}_t)\right)\right] \\
&\le \frac{2}{N}\left(\frac{1}{2\pi(1 - \bar{\alpha}_t)}\right)^{\frac{d}{2}}.
\end{aligned}
\tag{68}
$$

Plugging this into the above equation, we get

$$\mathbb{E}_{\boldsymbol{S}}\left[\exp\left(\lambda\left[f_{\boldsymbol{S}}(\boldsymbol{x}_t) - P_t(\boldsymbol{x}_t)\right]\right)\right] \le \exp\left(\frac{\lambda^2}{2n\left(2\pi(1 - \bar{\alpha}_t)\right)^d}\right), \tag{69}$$

which shows that $f_{\boldsymbol{S}}(\boldsymbol{x}_t) - P_t(\boldsymbol{x}_t)$ is a sub-Gaussian process w.r.t. $\boldsymbol{x}_t$.

Due to $\boldsymbol{x}_t$ has bounded norm, there exists a δ-cover $\mathcal{C}(\delta, D)$ of l_2-ball with radius D such that for any $\boldsymbol{x}_t$ there exists $\boldsymbol{y}_t \in \mathcal{C}(\delta, D)$ with $\|\boldsymbol{x}_t - \boldsymbol{y}_t\| \le \delta$. Due to Lemma 3

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{S}}\left[\sup_{\boldsymbol{x}_t}(f_{\boldsymbol{S}}(\boldsymbol{x}_t) - P_t(\boldsymbol{x}_t))\right] &= \mathbb{E}_{\boldsymbol{S}}\left[\sup_{\substack{\boldsymbol{x}_t, \boldsymbol{y}_t; \\ \|\boldsymbol{x}_t - \boldsymbol{y}_t\| \le \delta}} (f_{\boldsymbol{S}}(\boldsymbol{x}_t) - P_t(\boldsymbol{x}_t)) - (f_{\boldsymbol{S}}(\boldsymbol{y}_t) - P_t(\boldsymbol{y}_t))\right] \\
&\quad + \mathbb{E}\left[\sup_{\boldsymbol{y}_t \in \mathcal{C}(\delta, D)} (f_{\boldsymbol{S}}(\boldsymbol{y}_t) - P_t(\boldsymbol{y}_t))\right] \\
&\le 2\delta\sqrt{\frac{1}{e(1 - \bar{\alpha}_t)}} + \sqrt{\frac{2\log|\mathcal{C}(\delta, D)|}{n\left(2\pi(1 - \bar{\alpha}_t)\right)^d}},
\end{aligned}
\tag{70}
$$

where the last inequality is due to (69) and Exercise 3.7 in (Duchi, 2016). Due to the arbitrarity of δ and taking $n \to \infty$, we get $\mathbb{E}_{\boldsymbol{S}}\left[\sup_{\boldsymbol{x}_t}(f_{\boldsymbol{S}}(\boldsymbol{x}_t) - P_t(\boldsymbol{x}_t))\right] \longrightarrow 0$, which implies $f_{\boldsymbol{S}}(\boldsymbol{x}_t) \xrightarrow{P} P_t(\boldsymbol{x}_t$ for any $\boldsymbol{x}_t$.

Thus we show that denominator of (13) converge to the one of (60) in probability. Similarly, we can prove the numerator of (13) converge to the one of (60) in probability. First, we have

$$\mathbb{E}_{\boldsymbol{S}}\left[\nabla_{\boldsymbol{x}} f_{\boldsymbol{S}}(\boldsymbol{x}_t)\right] = \mathbb{E}_{\boldsymbol{x}_0}\left[\nabla_{\boldsymbol{x}} P_{t|0}(\boldsymbol{x}_t \mid \boldsymbol{x}_0)\right] = \nabla_{\boldsymbol{x}} P_t(\boldsymbol{x}_t). \tag{71}$$

Then,

$$\sup_{\boldsymbol{x}_t:\|\boldsymbol{x}_t\|<D} \|\nabla_{\boldsymbol{x}} f_{\boldsymbol{S}}(\boldsymbol{x}_t) - \nabla_{\boldsymbol{x}} P_t(\boldsymbol{x}_t)\| - \sup_{\boldsymbol{x}_t:\|\boldsymbol{x}_t\|<D} \|\nabla_{\boldsymbol{x}} f_{\boldsymbol{S}^{i'}}(\boldsymbol{x}_t) - \nabla_{\boldsymbol{x}} P_t(\boldsymbol{x}_t)\|$$

$$\leq \sup_{\boldsymbol{x}_t:\|\boldsymbol{x}_t\|<D} \|\nabla_{\boldsymbol{x}} f_{\boldsymbol{S}}(\boldsymbol{x}_t) - \nabla_{\boldsymbol{x}} f_{\boldsymbol{S}^{i'}}(\boldsymbol{x}_t)\|$$

$$= \frac{1}{n}\left(\frac{1}{2\pi(1-\bar{\alpha}_t)}\right)^{\frac{d}{2}} \sup_{\boldsymbol{x}_t}\left\| K_t(\boldsymbol{x}_t,\boldsymbol{x}_0^i)\left(\frac{\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^i}{1-\bar{\alpha}_t}\right) - K_t(\boldsymbol{x}_t,\boldsymbol{x}_0^{i'})\left(\frac{\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^{i'}}{1-\bar{\alpha}_t}\right)\right\| \quad (72)$$

$$\leq \frac{1}{n}\left(\frac{1}{2\pi(1-\bar{\alpha}_t)}\right)^{\frac{d}{2}} \sup_{\boldsymbol{x}_t,\boldsymbol{x}_0} K_t(\boldsymbol{x}_t,\boldsymbol{x}_0)\left\|\frac{\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^{i'}}{1-\bar{\alpha}_t}\right\|$$

$$\leq \frac{1}{n}\left(\frac{1}{2\pi(1-\bar{\alpha}_t)}\right)^{\frac{d}{2}} \sqrt{\frac{1}{e(1-\bar{\alpha}_t)}},$$

where the last inequality is due to $axe^{-\frac{ax^2}{2}} \leq \sqrt{a/e}$. Thus by McDiarmid's inequality, we must have

$$\mathbb{P}\left(\left|\sup_{\boldsymbol{x}_t:\|\boldsymbol{x}_t\|\leq D} \|\nabla_{\boldsymbol{x}} f_{\boldsymbol{S}}(\boldsymbol{x}_t) - \nabla_{\boldsymbol{x}} P_t(\boldsymbol{x}_t)\| - \mathbb{E}\left[\sup_{\boldsymbol{x}_t:\|\boldsymbol{x}_t\|\leq D} \|\nabla_{\boldsymbol{x}} f_{\boldsymbol{S}}(\boldsymbol{x}_t) - \nabla_{\boldsymbol{x}} P_t(\boldsymbol{x}_t)\|\right]\right| \geq \epsilon\right)$$

$$\leq \exp\left(-2en(2\pi(1-\bar{\alpha}_t))^d(1-\bar{\alpha}_t)\epsilon^2\right). \quad (73)$$

Then we show that $\sup_{\boldsymbol{x}_t:\|\boldsymbol{x}_t\|\leq D}\|\nabla_{\boldsymbol{x}} f_{\boldsymbol{S}}(\boldsymbol{x}_t) - \nabla_{\boldsymbol{x}} P_t(\boldsymbol{x}_t)\|$ converge to its expectation in probability. What left is showing its expectation converges to zero. Similar to the proof of (69), we can prove

$$\mathbb{E}_{\boldsymbol{S}}\left[\exp\left(\lambda\left[\|\nabla_{\boldsymbol{x}_t} f_{\boldsymbol{S}}(\boldsymbol{x}_t) - \nabla_{\boldsymbol{x}_t} P_t(\boldsymbol{x}_t)\|\right]\right)\right] \leq \exp\left(\sum_{j=1}^n \frac{\lambda^2\|\nabla_{\boldsymbol{x}_t} U_j - \nabla_{\boldsymbol{x}_t} L_j\|^2}{8}\right). \quad (74)$$

On the other hand, due to Lemma 3, we have

$$\|\nabla_{\boldsymbol{x}_t} U_j - \nabla_{\boldsymbol{x}_t} L_j\| \leq \sup_{\boldsymbol{x}_0^i,\boldsymbol{x}_0^{i'}} \frac{1}{n}\left\|\left[\left(\nabla_{\boldsymbol{x}_t} f_{\boldsymbol{x}_0^i}(\boldsymbol{x}_t) - \nabla_{\boldsymbol{x}_t} P_t(\boldsymbol{x}_t)\right) - \left(\nabla_{\boldsymbol{x}_t} f_{\boldsymbol{x}_0^{i'}}(\boldsymbol{x}_t) - \nabla_{\boldsymbol{x}_t} P_t(\boldsymbol{x}_t)\right)\right]\right\|$$

$$\leq \frac{4}{N}\left(\frac{1}{2\pi(1-\bar{\alpha}_t)}\right)^{\frac{d}{2}} \sqrt{\frac{1}{e(1-\bar{\alpha}_t)}}, \quad (75)$$

which implies

$$\mathbb{E}_{\boldsymbol{S}}\left[\exp\left(\lambda\left[\|\nabla_{\boldsymbol{x}_t} f_{\boldsymbol{S}}(\boldsymbol{x}_t) - \nabla_{\boldsymbol{x}_t} P_t(\boldsymbol{x}_t)\|\right]\right)\right] \leq \exp\left(\frac{2\lambda^2}{en(2\pi(1-\bar{\alpha}_t))^{\frac{d}{2}}(1-\bar{\alpha}_t)}\right). \quad (76)$$

Thus, due to Lemma 3,

$$\mathbb{E}_{\boldsymbol{S}}\left[\sup_{\boldsymbol{x}_t} \|\nabla_{\boldsymbol{x}_t} f_{\boldsymbol{S}}(\boldsymbol{x}_t) - \nabla_{\boldsymbol{x}_t} P_t(\boldsymbol{x}_t)\|\right]$$

$$\leq \mathbb{E}_{\boldsymbol{S}}\left[\sup_{\substack{\boldsymbol{x}_t,\boldsymbol{y}_t;\\\|\boldsymbol{x}_t-\boldsymbol{y}_t\|\leq\delta}} \|(\nabla_{\boldsymbol{x}_t} f_{\boldsymbol{S}}(\boldsymbol{x}_t) - \nabla_{\boldsymbol{x}_t} P_t(\boldsymbol{x}_t)) - (\nabla_{\boldsymbol{y}_t} f_{\boldsymbol{S}}(\boldsymbol{y}_t) - \nabla_{\boldsymbol{y}_t} P_t(\boldsymbol{y}_t))\|\right]$$

$$+ \mathbb{E}\left[\sup_{\boldsymbol{y}_t\in\mathcal{C}(\delta,D)} \|\nabla_{\boldsymbol{y}_t} f_{\boldsymbol{S}}(\boldsymbol{y}_t) - \nabla_{\boldsymbol{y}_t} P_t(\boldsymbol{y}_t)\|\right] \quad (77)$$

$$\leq 2\delta\left(\frac{1}{2\pi(1-\bar{\alpha}_t)}\right)^{\frac{d}{2}}\left(\frac{2+e}{e(1-\bar{\alpha}_t)}\right) + \sqrt{\frac{8\log|\mathcal{C}(\delta,D)|}{en(2\pi(1-\bar{\alpha}_t))^d(1-\bar{\alpha}_t)}},$$

By taking a proper δ and $n \to \infty$, we show that $\mathbb{E}_{\boldsymbol{S}}\left[\sup_{\boldsymbol{x}_t} \|\nabla_{\boldsymbol{x}_t} f_{\boldsymbol{S}}(\boldsymbol{x}_t) - \nabla_{\boldsymbol{x}_t} P_t(\boldsymbol{x}_t)\|\right]$ converges to zero. Thus, the denominator and numerator of (13) are respectively converge to the ones of (60).

Finally, due to $\|\boldsymbol{x}_t\|$ is bounded, $P_t(\boldsymbol{x})$ and $\nabla_{\boldsymbol{x}_t} P_t(\boldsymbol{x}_t)$ are all continuous, $\nabla_{\boldsymbol{x}_t} \log P_t(\boldsymbol{x}_t)$ is continuous. Thus by Slutsky's theorem (Shiryaev, 2016), we prove our result. □

D Proofs in Section 6

Lemma 1. *For and $s < t$, we have* $\frac{\mathbb{E}[\boldsymbol{\xi}_{t,s}|\boldsymbol{x}_t]}{\sqrt{1-r_{t,s}}} = \frac{\mathbb{E}[\boldsymbol{\xi}_{t,t-1}|\boldsymbol{x}_t]}{\sqrt{1-r_{t,t-1}}} = -\nabla_{\boldsymbol{x}_t} \log P_t(\boldsymbol{x}_t)$.

Proof. Due to (18) and Tweedie's formula (Efron, 2011), we know

$$\mathbb{E}[\boldsymbol{\xi}_{t,s} \mid \boldsymbol{x}_t] = \frac{1}{\sqrt{1-r_{t,s}}}\boldsymbol{x}_t - \frac{1}{\sqrt{1-r_{t,s}}}\left(\boldsymbol{x}_t + (1-r_{t,s})\nabla_{\boldsymbol{x}}\log P_t(\boldsymbol{x}_t)\right) = -\sqrt{1-r_{t,s}}\nabla_{\boldsymbol{x}}\log P_t(\boldsymbol{x}_t).$$
(78)

Thus $\mathbb{E}[\boldsymbol{\xi}_{t,s} \mid \boldsymbol{x}_t]/\sqrt{1-r_{t,s}}$ is invariant w.r.t. s, which verifies our conclusion. $\square$

Proposition 5. *Suppose the model $\boldsymbol{\xi}_{\boldsymbol{\theta}}(\cdot, \cdot)$ has enough functional capacity the optimal solution of* (21) *is*

$$\boldsymbol{\xi}_{\boldsymbol{\theta}_S}^*(\boldsymbol{x}, t) = \sum_{i=1}^n \frac{\mathbb{E}_s\left[\left(\frac{1}{2\pi(1-r_{t,s})}\right)^{\frac{d}{2}} \exp\left(-\frac{\|\boldsymbol{x}-\sqrt{r_{t,s}}\boldsymbol{x}_s^i\|^2}{2(1-r_{t,s})}\right)\left(\frac{\boldsymbol{x}-\sqrt{r_{t,s}}\boldsymbol{x}_s^i}{1-r_{t,s}}\right)\right]}{\sum_{i=1}^n \mathbb{E}_s\left[\left(\frac{1}{2\pi(1-r_{t,s})}\right)^{\frac{d}{2}} \exp\left(-\frac{\|\boldsymbol{x}-\sqrt{r_{t,s}}\boldsymbol{x}_s^i\|^2}{2(1-r_{t,s})}\right)\right]}.$$
(22)

Proof. Due to (18), for any t, our training objective (21) can be written as

$$\inf_{\boldsymbol{\xi}_{\boldsymbol{\theta}}} \mathbb{E}_s\left[\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{\xi}_{t,s}}\left[\left\|\frac{\boldsymbol{\xi}_{t,s}}{\sqrt{1-r_{t,s}}} - \boldsymbol{\xi}_{\boldsymbol{\theta}}(\sqrt{r_{t,s}}\boldsymbol{x}_s^i + \sqrt{1-r_{t,s}}\boldsymbol{\xi}_{t,s}, t)\right\|^2\right]\right]$$

$$= \inf_{\boldsymbol{\xi}_{\boldsymbol{\theta}}} \mathbb{E}_s\left[\frac{1}{n}\sum_{i=1}^n \int_{\mathbb{R}^d} \left\|\frac{\boldsymbol{\xi}_{t,s}}{\sqrt{1-r_{t,s}}} - \boldsymbol{\xi}_{\boldsymbol{\theta}}(\sqrt{r_{t,s}}\boldsymbol{x}_s^i + \sqrt{1-r_{t,s}}\boldsymbol{\xi}_{t,s}, t)\right\|^2 \left(\frac{1}{2\pi}\right)^{\frac{d}{2}}\exp\left(\frac{-\|\boldsymbol{\xi}_{t,s}\|^2}{2}\right)d\boldsymbol{\xi}_{t,s}\right]$$

$$= \inf_{\boldsymbol{\xi}_{\boldsymbol{\theta}}} \mathbb{E}_s\left[\frac{1}{n}\sum_{i=1}^n \int_{\mathbb{R}^d} \left\|\boldsymbol{\xi}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \frac{\boldsymbol{x}-\sqrt{r_{t,s}}\boldsymbol{x}_s^i}{1-r_{t,s}}\right\|^2 \left(\frac{1}{2\pi(1-r_{t,s})}\right)^{\frac{d}{2}}\exp\left(-\frac{\|\boldsymbol{x}-\sqrt{r_{t,s}}\boldsymbol{x}_s^i\|^2}{2(1-r_{t,s})}\right)d\boldsymbol{x}\right].$$
(79)

Then following the proof of Theorem 2, we prove our conclusion. $\square$

As we have clarified in the mainbody of this paper, the objective (20) has global minima $\mathbb{E}[\boldsymbol{\xi}_{t,t-1} \mid \boldsymbol{x}_t]$. We formally prove this conclusion in the following lemma.

Lemma 5. *For $\boldsymbol{\xi}_{\boldsymbol{\theta}}(\cdot, \cdot)$ with enough functional capacity, the problem* (20) *has global minima $\boldsymbol{\xi}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) = \mathbb{E}[\boldsymbol{\xi}_{t,t-1} \mid \boldsymbol{x}_t]/\sqrt{1-r_{t,t-1}}$.*

Proof. For any specific t, due to the optimality of conditional expectation of minimizing min-square estimation (Banerjee et al., 2005),

$$\inf_{\boldsymbol{\xi}_{\boldsymbol{\theta}}(\cdot, t)} \mathbb{E}_s\left[\mathbb{E}_{\boldsymbol{x}_s, \boldsymbol{\xi}_{t,s}}\left[\left\|\frac{\boldsymbol{\xi}_{t,s}}{\sqrt{1-r_{t,s}}} - \boldsymbol{\xi}_{\boldsymbol{\theta}}(\sqrt{r_{t,s}}\boldsymbol{x}_s + \sqrt{1-r_{t,s}}\boldsymbol{\xi}_{t,s}, t)\right\|^2\right]\right]$$

$$\geq \mathbb{E}_s\left[\inf_{\boldsymbol{\xi}_{\boldsymbol{\theta}}(\cdot, t)} \mathbb{E}_{\boldsymbol{x}_s, \boldsymbol{\xi}_{t,s}}\left[\left\|\frac{\boldsymbol{\xi}_{t,s}}{\sqrt{1-r_{t,s}}} - \boldsymbol{\xi}_{\boldsymbol{\theta}}(\sqrt{r_{t,s}}\boldsymbol{x}_s + \sqrt{1-r_{t,s}}\boldsymbol{\xi}_{t,s}, t)\right\|^2\right]\right]$$
(80)

$$= \mathbb{E}_s\left[\mathbb{E}_{\boldsymbol{x}_s, \boldsymbol{\xi}_{t,s}}\left[\left\|\frac{\boldsymbol{\xi}_{t,s}}{\sqrt{1-r_{t,s}}} - \mathbb{E}\left[\frac{\boldsymbol{\xi}_{t,s}}{\sqrt{1-r_{t,s}}} \mid \boldsymbol{x}_t\right]\right\|^2\right]\right],$$

where the first inequality becomes when $\boldsymbol{\xi}_{\boldsymbol{\theta}}(\sqrt{r_{t,s}}\boldsymbol{x}_s + \sqrt{1-r_{t,s}}\boldsymbol{\xi}_{t,s}, t) = \mathbb{E}\left[\frac{\boldsymbol{\xi}_{t,s}}{\sqrt{1-r_{t,s}}} \mid \boldsymbol{x}_t\right]$, which is invariant w.r.t. s due to Lemma 1. $\square$

It worth noting that our training objective is another view of score matching (Song et al., 2020), which approximate score function $\nabla_{\boldsymbol{x}_t} \log P_t(\boldsymbol{x}_t)$. Then using the approximated $\nabla_{\boldsymbol{x}_t} \log P_t(\boldsymbol{x}_t)$ to running a reverse-time stochastic differential equation to generate data. In this regime, they leverage a model $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ to minimizing $\mathbb{E}_{\boldsymbol{x}_t}[\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \nabla_{\boldsymbol{x}_t} \log P_t(\boldsymbol{x}_t)\|^2]$ to get the approximated score function $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$. It has been proven in (Vincent, 2011) that for any $s < t$, it holds

$$\inf_{\boldsymbol{s}_{\boldsymbol{\theta}}} \mathbb{E}_{\boldsymbol{x}_t}\left[\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \nabla_{\boldsymbol{x}_t} \log P_t(\boldsymbol{x}_t)\|^2\right] = \inf_{\boldsymbol{s}_{\boldsymbol{\theta}}} \mathbb{E}_{\boldsymbol{x}_s}\left[\mathbb{E}_{\boldsymbol{x}_t|\boldsymbol{x}_s}\left[\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \nabla_{\boldsymbol{x}_t} \log P_{t|s}(\boldsymbol{x}_t \mid \boldsymbol{x}_s)\|^2\right]\right].$$
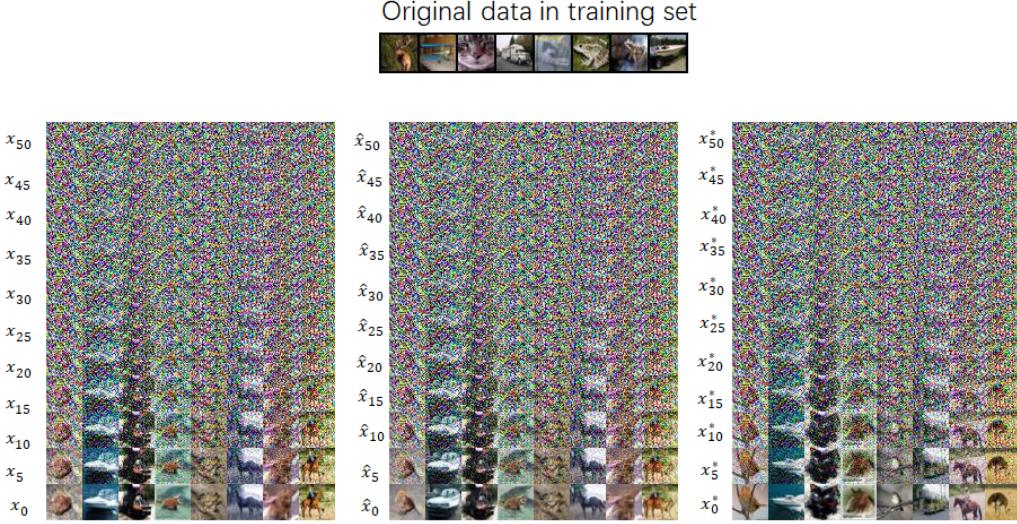(81)

Figure 4: The generated `CIFAR10`, starting with noisy data constructed by training set. From the left to right are respectively the data generated by diffusion models trained by (21) and (12).

Due to (18), we know that our training objective (20) is equivalent to

$$\inf_{\boldsymbol{s_\theta}} \mathbb{E}_s \left[\mathbb{E}_{\boldsymbol{x}_s} \left[\mathbb{E}_{\boldsymbol{x}_t | \boldsymbol{x}_s} \left[\left\| \boldsymbol{s_\theta}(\boldsymbol{x}_t, t) - \nabla_{\boldsymbol{x}_t} \log P_{t|s}(\boldsymbol{x}_t \mid \boldsymbol{x}_s) \right\|^2 \right] \right] \right], \tag{82}$$

which is equivalent to score-matching as in (Song et al., 2020) but with a random initial time step s (the s in (Song et al., 2020) is fixed as zero).

Theorem 4. *Let $\boldsymbol{\xi}_{\boldsymbol{\theta}_S}(\cdot, \cdot)$ be the model defined in (22), then for any t and $\boldsymbol{x}_t$ with bounded norm, we have $\boldsymbol{\xi}^*_{\boldsymbol{\theta}_S}(\boldsymbol{x}_t, t) \xrightarrow{P} \mathbb{E}[\boldsymbol{\xi}_{t,t-1} \mid \boldsymbol{x}_t] / \sqrt{1 - r_{t,t-1}}.$*

Proof. For any given t and $s < t$, we know

$$P_{t|s}(\boldsymbol{x}_t \mid \boldsymbol{x}_s) = \left(\frac{1}{2\pi(1 - r_{t,s})} \right)^{\frac{d}{2}} \exp \left(-\frac{\| \boldsymbol{x} - \sqrt{r_{t,s}} \boldsymbol{x}_s \|^2}{2(1 - r_{t,s})} \right);$$

$$\nabla_{\boldsymbol{x}_t} P_{t|s}(\boldsymbol{x}_t \mid \boldsymbol{x}_s) = \left(\frac{1}{2\pi(1 - r_{t,s})} \right)^{\frac{d}{2}} \exp \left(-\frac{\| \boldsymbol{x} - \sqrt{r_{t,s}} \boldsymbol{x}_s^i \|^2}{2(1 - r_{t,s})} \right) \left(\frac{\boldsymbol{x} - \sqrt{r_{t,s}} \boldsymbol{x}_s}{1 - r_{t,s}} \right). \tag{83}$$

Thus, the numerator and denominator are respectively unbiased estimator to $\nabla_{\boldsymbol{x}_t} P_t(\boldsymbol{x}_t)$ and $P_t(\boldsymbol{x}_t)$. Since s is finite, and $\mathbb{E}[\boldsymbol{\xi}_{t,t-1} \mid \boldsymbol{x}_t] / \sqrt{1 - r_{t,t-1}} = \nabla \log P_t(\boldsymbol{x}_t)$, we can similarly prove our result as in Theorem 3. ∎

E   Extra Experiments

In this section, we present some of generated data by different diffusion models in the main part of this paper.

E.1   Accumulated Optimization Bias Improves Generalization

In section 5.2, we have verified that the generated $\boldsymbol{x}_0$ can exist in training set when starting from $\boldsymbol{x}_{15}$ generated by data in training set. However, we claim that when accumulating enough bias during the reverse process of generating, the generalization problem can be obviated. That says we start the reverse process from the same $\boldsymbol{x}_{50}$ generated by the data in training set. The results are in Figure 4 and 5. As can be seen, the generated data do not visually similar to the original training data.

Finally, as we have claim in Section 5.2, the optimization bias enables the trained diffusion model to obviate generate data existed in the training set. For each generated data, we verifies it by searching the nearest data in the training set. Some of generated data are in Figure 6 and 7.
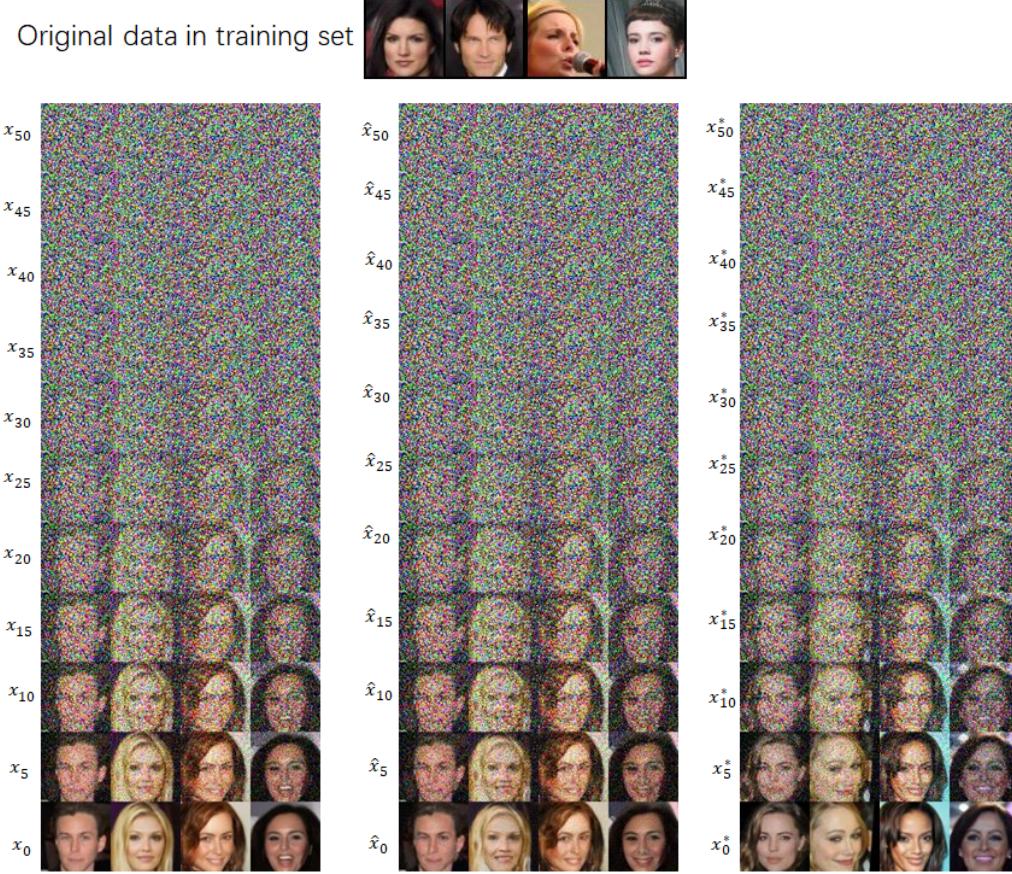
Figure 5: The generated `CelebA`, starting with noisy data constructed by training set. From the left to right are respectively the data generated by diffusion models trained by (21) and (12).

E.2 Data Generated by Different Diffusion Models

We present batch of generated data by different diffusion models, i.e., the ones trained by (12), (21) and the empirical optima. They are respectively represented by $\boldsymbol{x}_t$, $\hat{\boldsymbol{x}}_t$, and $\boldsymbol{x}_t^*$, and staring with the same standard Gaussian noise. Similar to Section 5, the data are generated by 50 steps DDIM (Song et al., 2022). The `CIFAR10` and `CelebA` are respectively in Figure 9 and 8. As can be seen, the $\boldsymbol{x}_t$ and $\hat{\boldsymbol{x}}_t$ are close to each other, while $\hat{\boldsymbol{x}}_t$ is noisy than $\boldsymbol{x}_t$. This further verifies there is a trade-off between generalization and optimization as we discussed in the Section 7.

E.3 Generated $\hat{\boldsymbol{x}}_0$

In this subsection, we compare some data generated by the diffusion model trained by (21) and (12). Though the first model has no potential generalization problem, its generated data are noisy compared with $\boldsymbol{x}_t$. The data are in Figure 10 and 11.

The nearest data in the training set

x_0 $\hat{x}_0$ x_0^*

Figure 6: The generated CIFAR10, the bottom and top line are respectively the generated data and the its nearest data in the training set. From the left to right are respectively the data generated by diffusion models trained by (12), (12) and empirical optima.
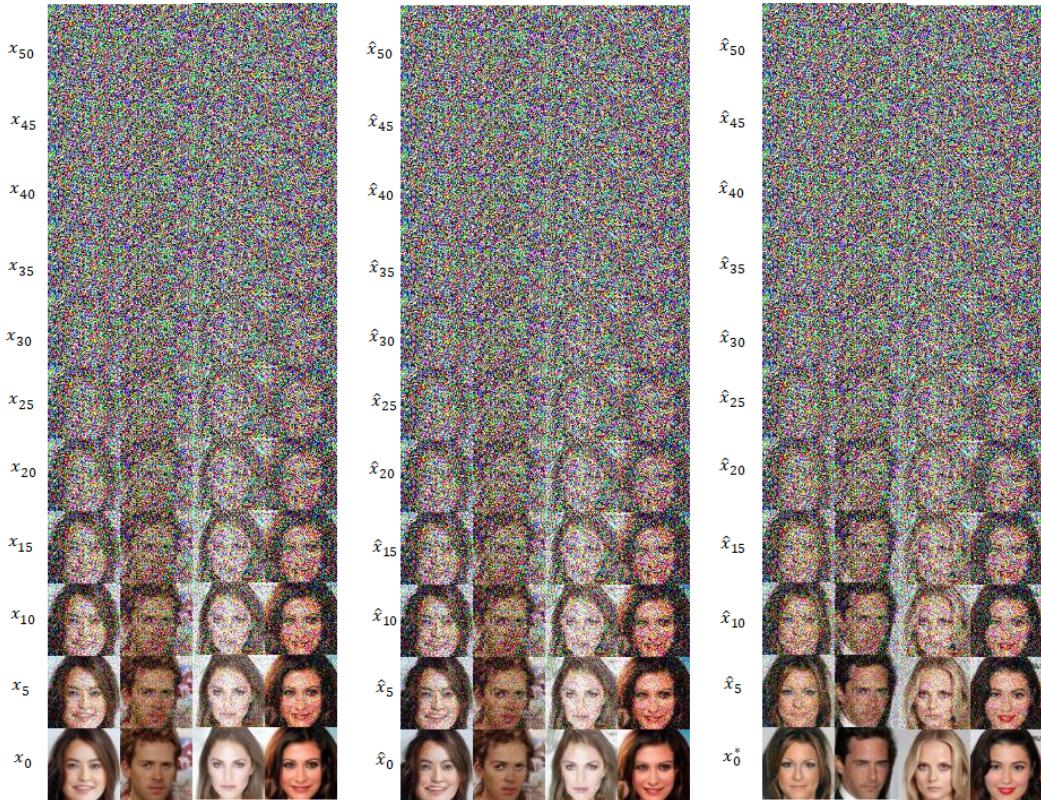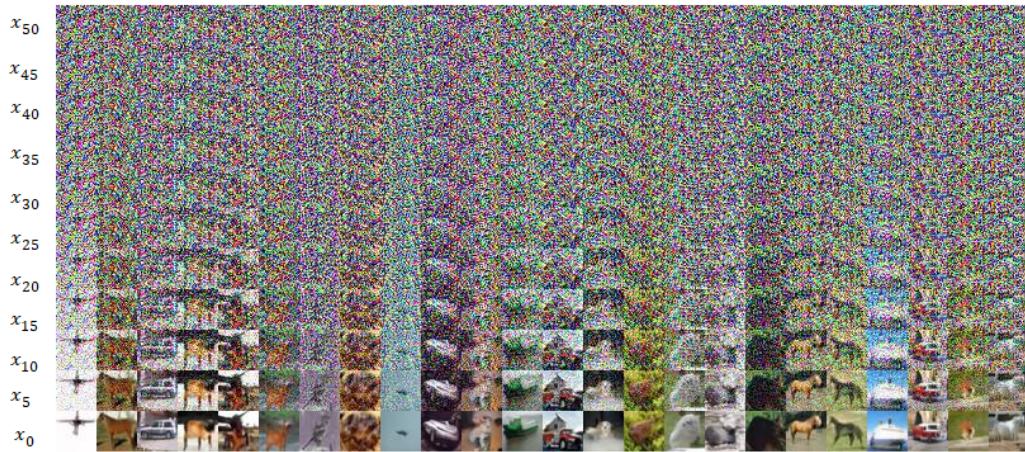
The nearest data in the training set

x_0 $\hat{x}_0$ x_0^*

Figure 7: The generated CelebA, the bottom and top line are respectively the generated data and the its nearest data in the training set. From the left to right are respectively the data generated by diffusion models trained by (12), (12) and empirical optima.
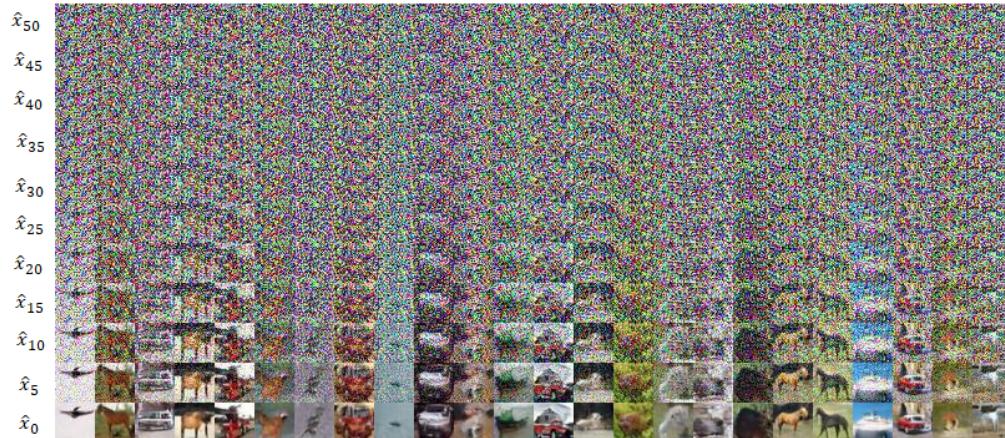
Figure 8: The generated `CelebA`, from the left to right are respectively the data generated by diffusion models trained by (12), (21) and the empirical optima (13).

(a) CIFAR10 generated by the model trained by (12).

(b) CIFAR10 generated by the model trained by (22).

(c) CIFAR10 generated by the model trained empirical optima (13).

Figure 9: The generated CIFAR10

Figure 10: The generated `CIFAR10`, from the left to right are respectively the data generated by diffusion models trained by (21) and (12).

Figure 11: The generated `CelebA`, from the left to right are respectively the data generated by diffusion models trained by (21) and (12).