Examining the Convergence of Denoising Diffusion Probabilistic Models: A Quantitative Analysis

Abstract

Deep generative models, particularly diffusion models, are a significant family within deep learning. This study provides a precise upper limit for the Wasserstein distance between a learned distribution by a diffusion model and the target distribution. In contrast to earlier research, this analysis does not rely on presumptions regarding the learned score function. Furthermore, the findings are applicable to any data-generating distributions within restricted instance spaces, even those lacking a density relative to the Lebesgue measure, and the upper limit is not exponentially dependent on the ambient space dimension. The primary finding expands upon recent research by Mbacke et al. (2023), and the proofs presented are fundamental.

1 Introduction

Diffusion models, alongside generative adversarial networks and variational autoencoders (VAEs), are among the most influential families of deep generative models. These models have demonstrated remarkable empirical results in generating images and audio, as well as in various other applications.

Two primary methods exist for diffusion models: denoising diffusion probabilistic models (DDPMs) and score-based generative models (SGMs). DDPMs incrementally convert samples from the desired distribution into noise via a forward process, while simultaneously training a backward process to reverse this transformation, enabling the creation of new samples. Conversely, SGMs employ score-matching methods to approximate the score function of the data-generating distribution, subsequently generating new samples through Langevin dynamics. Recognizing that real-world distributions might lack a defined score function, adding varying noise levels to training samples to encompass the entire instance space and training a neural network to concurrently learn the score function for all noise levels has been proposed.

Although DDPMs and SGMs may initially seem distinct, it has been demonstrated that DDPMs implicitly approximate the score function, with the sampling process resembling Langevin dynamics. Moreover, a unified perspective of both methods using stochastic differential equations (SDEs) has been derived. The SGM can be viewed as a discretization of Brownian motion, and the DDPM as a discretization of an Ornstein-Uhlenbeck process. Consequently, both DDPMs and SGMs are commonly referred to as SGMs in the literature. This explains why prior research investigating the theoretical aspects of diffusion models has adopted the score-based framework, necessitating assumptions about the effectiveness of the learned score function.

In this research, a different strategy is employed, applying methods created for VAEs to DDPMs, which can be viewed as hierarchical VAEs with fixed encoders. This method enables the derivation of quantitative, Wasserstein-based upper bounds without making assumptions about the data distribution or the learned score function, and with simple proofs that do not need the SDE toolkit. Furthermore, the bounds presented here do not involve any complex discretization steps, as the forward and backward processes are considered discrete-time from the beginning, rather than being viewed as discretizations of continuous-time processes.

1.1 Related Works

There has been an increasing amount of research aimed at providing theoretical findings on the convergence of SGMs. However, these studies frequently depend on restrictive assumptions regarding the data-generating distribution, produce non-quantitative upper bounds, or exhibit exponential dependencies on certain parameters. This work successfully circumvents all three of these limitations. Some bounds are based on very restrictive assumptions about the data-generating distribution, such as log-Sobolev inequalities, which are unrealistic for real-world data distributions. Furthermore, some studies establish upper bounds on the Kullback-Leibler (KL) divergence or the total variation (TV) distance between the data-generating distribution and the distribution learned by the diffusion model; however, unless strong assumptions are made about the support of the data-generating distribution, KL and TV reach their maximum values. Such assumptions arguably do not hold for real-world data-generating distributions, which are widely believed to satisfy the manifold hypothesis. Other work establishes conditions under which the support of the input distribution is equal to the support of the learned distribution, and generalizes the bound to all f-divergences. Assuming L2 accurate score

estimation, some establish Wasserstein distance upper bounds under weaker assumptions on the data-generating distribution, but their Wasserstein-based bounds are not quantitative. Quantitative Wasserstein distance upper bounds under the manifold hypothesis have been derived, but these bounds exhibit exponential dependencies on some of the problem parameters.

1.2 Our contributions

In this study, strong assumptions about the data-generating distribution are avoided, and a quantitative upper bound on the Wasserstein distance is established without exponential dependencies on problem parameters, including the ambient space dimension. Moreover, a common aspect of the aforementioned studies is that their bounds are contingent on the error of the score estimator. According to some, providing precise guarantees for the estimation of the score function is challenging, as it necessitates an understanding of the non-convex training dynamics of neural network optimization, which is currently beyond reach. Therefore, upper bounds are derived without making assumptions about the learned score function. Instead, the bound presented here is dependent on a reconstruction loss calculated over a finite independent and identically distributed (i.i.d.) sample. Intuitively, a loss function is defined, which quantifies the average Euclidean distance between a sample from the data-generating distribution and the reconstruction obtained by sampling noise and passing it through the backward process (parameterized by ŏ3b8). This method is inspired by previous work on VAEs.

This approach offers numerous benefits: it does not impose restrictive assumptions on the data-generating distribution, avoids exponential dependencies on the dimension, and provides a quantitative upper bound based on the Wasserstein distance. Furthermore, this method benefits from utilizing very straightforward and basic proofs.

2 Preliminaries

Throughout this paper, lowercase letters are used to represent both probability measures and their densities with respect to the Lebesgue measure, and variables are added in parentheses to enhance readability (e.g., $q(x_t|x_{t-1})$ to denote a time-dependent conditional distribution). An instance space X, which is a subset of R^D with the Euclidean distance as the underlying metric, and a target data-generating distribution $\mu \in M_1^+(X)$ are considered. Note that it is not assumed that μ has a density with respect to the Lebesgue measure. Additionally, $||\cdot||$ represents the Euclidean (L2) norm, and $E_{p(x)}$ is used as shorthand for $E_{x\sim p(x)}$. Given probability measures $p, q \in M_1^+(X)$ and a real number $k > 1$, the Wasserstein distance of order k is defined as (Villani, 2009):

$$W_k(p,q) = \inf_{\gamma\in\Gamma(p,q)} \left(\int_{X\times X} ||x-y||^k d\gamma(x,y)\right)^{1/k},$$

where $\Gamma(p,q)$ denotes the set of couplings of p and q, meaning the set of joint distributions on $X \times X$ with respective marginals p and q. The product measure $p \otimes q$ is referred to as the trivial coupling, and the Wasserstein distance of order 1 is simply referred to as the Wasserstein distance.

2.1 Denoising Diffusion Models

Instead of employing the SDE framework, diffusion models are presented using the DDPM formulation with discrete-time processes. A diffusion model consists of two discrete-time stochastic processes: a forward process and a backward process. Both processes are indexed by time $0 \le t \le T$, where the number of time steps T is a predetermined choice.

The forward process. The forward process transforms a data point $x_0 \sim \mu$ into a noise distribution $q(x_T|x_0)$ through a sequence of conditional distributions $q(x_t|x_{t-1})$ for $1 \le t \le T$. It is assumed that the forward process is defined such that for sufficiently large T, the distribution $q(x_T|x_0)$ is close to a simple noise distribution $p(x_T)$, which is referred to as the prior distribution. For instance, $p(x_T) = N(x_T; 0, I)$, the standard multivariate normal distribution, has been chosen in previous work.

The backward process. The backward process is a Markov process with parametric transition kernels. The objective of the backward process is to perform the reverse operation of the forward process: transforming noise samples into (approximate) samples from the distribution μ. Following previous work, it is assumed that the backward process is defined by Gaussian distributions $p_\theta(x_{t-1}|x_t)$ for $2 \le t \le T$ as

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; g_t^\theta(x_t), \sigma_t^2 I),$$

and

$$p_\theta(x_0|x_1) = g_1^\theta(x_1),$$

where the variance parameters $\sigma_t^2 \in R_{\ge 0}$ are defined by a fixed schedule, the mean functions $g_t^\theta : R^D \to R^D$ are learned using a neural network (with parameters θ) for $2 \le t \le T$, and $g_1^\theta : R^D \to X$ is a separate function dependent on σ_1. In practice, the same network has been used for the functions g_t^θ for $2 \le t \le T$, and a separate discrete decoder for g_1^θ.

Generating new samples from a trained diffusion model is accomplished by sampling $x_{t-1} \sim p_\theta(x_{t-1}|x_t)$ for $1 \leq t \leq T$, starting from a noise vector $x_T \sim p(x_T)$ sampled from the prior $p(x_T)$.

The following assumption is made regarding the backward process.

Assumption 1. It is assumed that for each $1 \leq t \leq T$, there exists a constant $K_t^\theta > 0$ such that for every $x_1, x_2 \in X$,

$$||g_t^\theta(x_1) - g_t^\theta(x_2)|| \leq K_t^\theta ||x_1 - x_2||.$$

In other words, g_t^θ is K_t^θ-Lipschitz continuous. This assumption is discussed in Remark 3.2.

2.2 Additional Definitions

The distribution $\pi_\theta(\cdot|x_0)$ is defined as

$$\pi_\theta(\cdot|x_0) = q(x_T|x_0)p_\theta(x_{T-1}|x_T)p_\theta(x_{T-2}|x_{T-1}) \dots p_\theta(x_1|x_2)p_\theta(\cdot|x_1).$$

Intuitively, for each $x_0 \in X$, $\pi_\theta(\cdot|x_0)$ represents the distribution on X obtained by reconstructing samples from $q(x_T|x_0)$ through the backward process. Another way to interpret this distribution is that for any function $f : X \to R$, the following equation holds:

$$E_{\pi_\theta(\hat{x}_0|x_0)}[f(\hat{x}_0)] = E_{q(x_T|x_0)}E_{p_\theta(x_{T-1}|x_T)} \dots E_{p_\theta(x_1|x_2)}E_{p_\theta(\hat{x}_0|x_1)}[f(\hat{x}_0)].$$

Given a finite set $S = \{x_0^1, \dots, x_0^n\}$ i.i.d. $\sim \mu$, the regenerated distribution is defined as the following mixture:

$$\mu_n^\theta = \frac{1}{n}\sum_{i=1}^{n} \pi_\theta(\cdot|x_0^i).$$

This definition is analogous to the empirical regenerated distribution defined for VAEs. The distribution on X learned by the diffusion model is denoted as $\pi_\theta(\cdot)$ and defined as

$$\pi_\theta(\cdot) = p(x_T)p_\theta(x_{T-1}|x_T)p_\theta(x_{T-2}|x_{T-1}) \dots p_\theta(x_1|x_2)p_\theta(\cdot|x_1).$$

In other words, for any function $f : X \to R$, the expectation of f with respect to $\pi_\theta(\cdot)$ is

$$E_{\pi_\theta(\hat{x}_0)}[f(\hat{x}_0)] = E_{p(x_T)}E_{p_\theta(x_{T-1}|x_T)} \dots E_{p_\theta(x_1|x_2)}E_{p_\theta(\hat{x}_0|x_1)}[f(\hat{x}_0)].$$

Hence, both $\pi_\theta(\cdot)$ and $\pi_\theta(\cdot|x_0)$ are defined using the backward process, with the difference that $\pi_\theta(\cdot)$ starts with the prior $p(x_T) = N(x_T; 0, I)$, while $\pi_\theta(\cdot|x_0)$ starts with the noise distribution $q(x_T|x_0)$.

Finally, the loss function $l_\theta : X \times X \to R$ is defined as

$$l_\theta(x_T, x_0) = E_{p_\theta(x_{T-1}|x_T)}E_{p_\theta(x_{T-2}|x_{T-1})} \dots E_{p_\theta(x_1|x_2)}E_{p_\theta(\hat{x}_0|x_1)}[||x_0 - \hat{x}_0||].$$

Hence, given a noise vector x_T and a sample x_0, the loss $l_\theta(x_T, x_0)$ represents the average Euclidean distance between x_0 and any sample obtained by passing x_T through the backward process.

2.3 Our Approach

The goal is to upper-bound the distance $W_1(\mu, \pi_\theta(\cdot))$. Since the triangle inequality implies

$$W_1(\mu, \pi_\theta(\cdot)) \leq W_1(\mu, \mu_n^\theta) + W_1(\mu_n^\theta, \pi_\theta(\cdot)),$$

the distance $W_1(\mu, \pi_\theta(\cdot))$ can be upper-bounded by upper-bounding the two expressions on the right-hand side separately. The upper bound on $W_1(\mu, \mu_n^\theta)$ is obtained using a straightforward adaptation of a proof. First, $W_1(\mu, \mu_n^\theta)$ is upper-bounded using the expectation of the loss function l_θ, then the resulting expression is upper-bounded using a PAC-Bayesian-style expression dependent on the empirical risk and the prior-matching term.

The upper bound on the second term $W_1(\mu_n^\theta, \pi_\theta(\cdot))$ uses the definition of μ_n^θ. Intuitively, the difference between $\pi_\theta(\cdot|x_0^i)$ and $\pi_\theta(\cdot)$ is determined by the corresponding initial distributions: $q(x_T|x_0^i)$ and $p(x_T)$ for $\pi_\theta(\cdot)$. Hence, if the two initial distributions are close, and if the steps of the backward process are smooth (see Assumption 1), then $\pi_\theta(\cdot|x_0^i)$ and $\pi_\theta(\cdot)$ are close to each other.

3 Main Result

3.1 Theorem Statement

We are now ready to present the main result: a quantitative upper bound on the Wasserstein distance between the data-generating distribution μ and the learned distribution $\pi_\theta(\cdot)$.

Theorem 3.1. Assume the instance space X has finite diameter $\Delta = \sup_{x,x' \in X} ||x - x'|| < \infty$, and let $\lambda > 0$ and $\delta \in (0, 1)$ be real numbers. Using the definitions and assumptions of the previous section, the following inequality holds with probability at least $1 - \delta$ over the random draw of $S = \{x_0^1, \ldots, x_0^n\}$ i.i.d. $\sim \mu$:

$$\begin{aligned} W_1(\mu, \pi_\theta(\cdot)) \leq & \frac{1}{n} \sum_{i=1}^{n} E_{q(x_T|x_0^i)}[l_\theta(x_T, x_0^i)] + \frac{1}{\lambda n} \sum_{i=1}^{n} KL(q(x_T|x_0^i)||p(x_T)) + \frac{1}{\lambda n} \log \frac{n}{\delta} + \frac{\lambda \Delta^2}{8n} \\ & + \left(\prod_{t=1}^{T} K_t^\theta \right) E_{q(x_T|x_0^i)} E_{p(y_T)}[||x_T - y_T||] \\ & + \sum_{t=2}^{T} \left(\prod_{i=1}^{t-1} K_i^\theta \right) \sigma_t E_{\epsilon, \epsilon'}[||\epsilon - \epsilon'||], \end{aligned}$$

where $\epsilon, \epsilon' \sim N(0, I)$ are standard Gaussian vectors.

Remark 3.1. Before presenting the proof, let us discuss Theorem 3.1.

* Because the right-hand side of the equation depends on a quantity computed using a finite i.i.d. sample S, the bound holds with high probability with respect to the randomness of S. This is the price we pay for having a quantitative upper bound with no exponential dependencies on problem parameters and no assumptions on the data-generating distribution μ. * The first term of the right-hand side is the average reconstruction loss computed over the sample $S = \{x_0^1, \ldots, x_0^n\}$. Note that for each $1 \leq i \leq n$, the expectation of $l_\theta(x_T|x_0^i)$ is only computed with respect to the noise distribution $q(x_T|x_0^i)$ defined by x_0^i itself. Hence, this term measures how well a noise vector $x_T \sim q(x_T|x_0^i)$ recovers the original sample x_0^i using the backward process, and averages over the set $S = \{x_0^1, \ldots, x_0^n\}$. * If the Lipschitz constants satisfy $K_t^\theta < 1$ for all $1 \leq t \leq T$, then the larger T is, the smaller the upper bound gets. This is because the product of K_t^θ's then converges to 0. In Remark 3.2 below, we show that the assumption that $K_t^\theta < 1$ for all t is a quite reasonable one. * The hyperparameter λ controls the trade-off between the prior-matching (KL) term and the diameter term Δ^2. If $K_t^\theta < 1$ for all $1 \leq t \leq T$ and $T \to \infty$, then the convergence of the bound largely depends on the choice of λ. In that case, $\lambda \propto n^{1/2}$ leads to faster convergence, while $\lambda \propto n$ leads to slower convergence to a smaller quantity. This is because the bound stems from PAC-Bayesian theory, where this trade-off is common. * The last term of the equation does not depend on the sample size n. Hence, the upper bound given by Theorem 3.1 does not converge to 0 as $n \to \infty$. However, if the Lipschitz factors $(K_t^\theta)_{1 \leq t \leq T}$ are all less than 1, then this term can be very small, especially in low-dimensional spaces.

3.2 Proof of the main theorem

The following result is an adaptation of a previous result.

Lemma 3.2. Let $\lambda > 0$ and $\delta \in (0, 1)$ be real numbers. With probability at least $1 - \delta$ over the randomness of the sample $S = \{x_0^1, \ldots, x_0^n\}$ i.i.d. $\sim \mu$, the following holds:

$$W_1(\mu, \mu_n^\theta) \leq \frac{1}{n} \sum_{i=1}^{n} E_{q(x_T|x_0^i)}[l_\theta(x_T, x_0^i)] + \frac{1}{\lambda n} \sum_{i=1}^{n} KL(q(x_T|x_0^i)||p(x_T)) + \frac{1}{\lambda n} \log \frac{n}{\delta} + \frac{\lambda \Delta^2}{8n}.$$

The proof of this result is a straightforward adaptation of a previous proof.

Now, let us focus our attention on the second term of the right-hand side of the equation, namely $W_1(\mu_n^\theta, \pi_\theta(\cdot))$. This part is trickier than for VAEs, for which the generative model's distribution is simply a pushforward measure. Here, we have a non-deterministic sampling process with T steps.

Assumption 1 leads to the following lemma on the backward process.

Lemma 3.3. For any given $x_1, y_1 \in X$, we have

$$E_{p_\theta(x_0|x_1)} E_{p_\theta(y_0|y_1)}[||x_0 - y_0||] \leq K_1^\theta ||x_1 - y_1||.$$

Moreover, if $2 \leq t \leq T$, then for any given $x_t, y_t \in X$, we have

$$E_{p_\theta(x_{t-1}|x_t)}E_{p_\theta(y_{t-1}|y_t)}[||x_{t-1}-y_{t-1}||] \leq K_t^\theta||x_t-y_t|| + \sigma_t E_{\epsilon,\epsilon'}[||\epsilon-\epsilon'||],$$

where $\epsilon, \epsilon' \sim N(0, I)$, meaning $E_{\epsilon,\epsilon'}$ is a shorthand for $E_{\epsilon,\epsilon' \sim N(0,I)}$.

Proof. For the first part, let $x_1, y_1 \in X$. Since according to the equation $p_\theta(x_0|x_1) = \delta_{g_1^\theta(x_1)}(x_0)$ and $p_\theta(y_0|y_1) = \delta_{g_1^\theta(y_1)}(y_0)$, then

$$E_{p_\theta(x_0|x_1)}E_{p_\theta(y_0|y_1)}[||x_0-y_0||] = ||g_1^\theta(x_1) - g_1^\theta(y_1)|| \leq K_1^\theta||x_1-y_1||.$$

For the second part, let $2 \leq t \leq T$ and $x_t, y_t \in X$. Since $p_\theta(x_{t-1}|x_t) = N(x_{t-1}; g_t^\theta(x_t), \sigma_t^2 I)$, the reparameterization trick implies that sampling $x_{t-1} \sim p_\theta(x_{t-1}|x_t)$ is equivalent to setting

$$x_{t-1} = g_t^\theta(x_t) + \sigma_t\epsilon_t, \text{ with } \epsilon_t \sim N(0, I).$$

Using the above equation, the triangle inequality, and Assumption 1, we obtain

$$\begin{aligned}
&E_{p_\theta(x_{t-1}|x_t)}E_{p_\theta(y_{t-1}|y_t)}[||x_{t-1}-y_{t-1}||] \\
&= E_{\epsilon_t,\epsilon_t' \sim N(0,I)}[||g_t^\theta(x_t) + \sigma_t\epsilon_t - g_t^\theta(y_t) - \sigma_t\epsilon_t'||] \\
&\leq E_{\epsilon_t,\epsilon_t' \sim N(0,I)}[||g_t^\theta(x_t) - g_t^\theta(y_t)||] + \sigma_t E_{\epsilon_t,\epsilon_t' \sim N(0,I)}[||\epsilon_t - \epsilon_t'||] \\
&\leq K_t^\theta||x_t-y_t|| + \sigma_t E_{\epsilon,\epsilon'}[||\epsilon-\epsilon'||],
\end{aligned}$$

where $\epsilon, \epsilon' \sim N(0, I)$.

Next, we can use the inequalities of Lemma 3.3 to prove the following result.

Lemma 3.4. Let $T \geq 1$. The following inequality holds:

$$\begin{aligned}
&E_{p_\theta(x_{T-1}|x_T)}E_{p_\theta(y_{T-1}|y_T)}E_{p_\theta(x_{T-2}|x_{T-1})}E_{p_\theta(y_{T-2}|y_{T-1})}\cdots E_{p_\theta(x_0|x_1)}E_{p_\theta(y_0|y_1)}[||x_0-y_0||] \\
&\leq \left(\prod_{t=1}^T K_t^\theta\right)||x_T-y_T|| + \sum_{t=2}^T\left(\prod_{i=1}^{t-1}K_i^\theta\right)\sigma_t E_{\epsilon,\epsilon'}[||\epsilon-\epsilon'||],
\end{aligned}$$

where $\epsilon, \epsilon' \sim N(0, I)$.

Proof Idea. Lemma 3.4 is proven by induction using Lemma 3.3 in the induction step.

Using the two previous lemmas, we obtain the following upper bound on $W_1(\mu_n^\theta, \pi_\theta(\cdot))$.

Lemma 3.5. The following inequality holds:

$$W_1(\mu_n^\theta, \pi_\theta(\cdot)) \leq \frac{1}{n}\sum_{i=1}^n\left(\prod_{t=1}^T K_t^\theta\right)E_{q(x_T|x_0^i)}E_{p(y_T)}[||x_T-y_T||] + \sum_{t=2}^T\left(\prod_{i=1}^{t-1}K_i^\theta\right)\sigma_t E_{\epsilon,\epsilon'}[||\epsilon-\epsilon'||],$$

where $\epsilon, \epsilon' \sim N(0, I)$.

Proof. Using the definition of W_1, the trivial coupling, the definitions of μ_n^θ and $\pi_\theta(\cdot)$, and Lemma 3.4, we get the desired result.

Combining Lemmas 3.2 and 3.5 with the triangle inequality yields Theorem 3.1.

3.3 Special case using the forward process of Ho et al. (2020)

Theorem 3.1 establishes a general upper bound that holds for any forward process, as long as the backward process satisfies Assumption 1. In this section, we specialize the statement of the theorem to the particular case of the forward process defined in previous work.

Let $X \subseteq R^D$. The forward process is a Gauss-Markov process with transition densities defined as

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I),$$

where $\alpha_1, \ldots, \alpha_T$ is a fixed noise schedule such that $0 < \alpha_t < 1$ for all t. This definition implies that at each time step $1 \leq t \leq T$,

$$q(x_t|x_0) = N(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I), \text{ with } \bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i.$$

The optimization objective to train the backward process ensures that for each time step t, the distribution $p_\theta(x_{t-1}|x_t)$ remains close to the ground-truth distribution $q(x_{t-1}|x_t, x_0)$ given by

$$q(x_{t-1}|x_t, x_0) = N(x_{t-1}; \tilde{\mu}_t^q(x_t, x_0), \tilde{\sigma}_t^2 I),$$

where

$$\tilde{\mu}_t^q(x_t, x_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}x_0.$$

Now, we discuss Assumption 1 under these definitions.

Remark 3.2. We can get a glimpse at the range of K_t^θ for a trained DDPM by looking at the distribution $q(x_{t-1}|x_t, x_0)$, since $p_\theta(x_{t-1}|x_t)$ is optimized to be as close as possible to $q(x_{t-1}|x_t, x_0)$.

For a given $x_0 \sim \mu$, let us take a look at the Lipschitz norm of $x \mapsto \tilde{\mu}_t^q(x, x_0)$. Using the above equation, we have

$$\tilde{\mu}_t^q(x_t, x_0) - \tilde{\mu}_t^q(y_t, x_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}(x_t - y_t).$$

Hence, $x \mapsto \tilde{\mu}_t^q(x, x_0)$ is K_t'-Lipschitz continuous with

$$K_t' = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}.$$

Now, if $\alpha_t < 1$ for all $1 \leq t \leq T$, then we have $1 - \bar{\alpha}_t > 1 - \bar{\alpha}_{t-1}$, which implies $K_t' < 1$ for all $1 \leq t \leq T$.

Remark 3.2 shows that the Lipschitz norm of the mean function $\tilde{\mu}_t^q(\cdot, x_0)$ does not depend on x_0. Indeed, looking at the previous equation, we can see that for any initial x_0, the Lipschitz norm $K_t' = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}$ only depends on the noise schedule, not x_0 itself. Since $g_t^\theta(\cdot, x_0)$ is optimized to match $\tilde{\mu}_t^q(\cdot, x_0)$ for each x_0 in the training set, and all the functions $\tilde{\mu}_t^q(\cdot, x_0)$ have the same Lipschitz norm K_t', we believe it is reasonable to assume g_t^θ is Lipschitz continuous as well. This is the intuition behind Assumption 1.

The prior-matching term. With the definitions of this section, the prior matching term $KL(q(x_T|x_0)||p(x_T))$ has the following closed form:

$$KL(q(x_T|x_0)||p(x_T)) = \frac{1}{2}\left[-D\log(1-\bar{\alpha}_T) - D\bar{\alpha}_T + \bar{\alpha}_T||x_0||^2\right].$$

Upper-bounds on the average distance between Gaussian vectors. If ϵ, ϵ' are D-dimensional vectors sampled from $N(0, I)$, then

$$E_{\epsilon,\epsilon'}[||\epsilon - \epsilon'||] \leq \sqrt{2D}.$$

Moreover, since $q(x_T|x_0) = N(x_T; \sqrt{\bar{\alpha}_T}x_0, (1-\bar{\alpha}_T)I)$ and the prior $p(y_T) = N(y_T; 0, I)$,

$$E_{q(x_T|x_0)}E_{p(y_T)}[||x_T - y_T||] \leq \sqrt{\bar{\alpha}_T||x_0||^2 + (2-\bar{\alpha}_T)D}.$$

Special case of the main theorem. With the definitions of this section, the inequality of Theorem 3.1 implies that with probability at least $1-\delta$ over the randomness of $\{x_0^1, \ldots, x$