Stat 689 Final report: Minimax Estimation in Diffusion Models

Jiyoung Park
Department of Statistics
Texas A&M University
wldyddl5510@tamu.edu

Abstract

In this final report, I will review the minimax optimality of the diffusion model proposed in Oko et al. [2023]. The key idea lies in the minimax estimation theory in Besov space established in Donoho and Johnstone [1998], Giné and Nickl [2015], Suzuki [2019]. By combining extension of Besov space result in Wasserstein space [Niles-Weed and Berthet, 2022] and a novel approximation method, the paper shows the specific structure in the diffusion model leads to the minimax optimal bound. I will go over the sketch of proofs, and discuss strength, weakness, and possible extension of this work.

1 Background

1.1 Main problem

Generative models Let $\mathcal{P}(\mathcal{X})$ be a set of probability measures in $\mathcal{X}$. Generative models aims to solve the following problem: suppose one have $X_i \overset{i.i.d}{\sim} P^* \in \mathcal{P}(\mathcal{X})$ for $i = 1, \ldots, n$. Given these data, is there an algorithm generating more samples $X_i \sim P^*$?

Among multiple algorithms for generative models, how one can say the certain algorithm is better than the other? If generated samples from the certain algorithm are very close to samples in P^*, then such algorithm can be regarded as a good algorithm. To precisely establish this procedure, one can bring back the classical statistical decision theory. Since we only work with data without knowing the true distribution P^*, what a generative model reproduces is in fact $\widehat{X}_i$, an estimator of X_i. Therefore, a statistical procedure to decide which estimator to use can directly be applied to this problem as well.

Minimax theory A minimax theory is one way of evaluating the performance of the estimator. In a general statistical framework, one quantifies the performance of the estimator based on certain types of 'risk' that the estimator have; an estimator of lower risk is more desirable. For example, in a linear regression problem, one can argue an estimator with the minimum mean-squared-error (MSE) is an ideal estimator. In this case, the risk functional will be $R(\theta, \widehat{\theta}) := \mathbb{E}\left\|\theta - \widehat{\theta}\right\|^2$.

One way to measure the *complexity of the problem* is by measuring the *minimax risk*. That is, given a risk functional $R(\theta, \widehat{\theta})$, one can consider the following quantity:

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} R(\theta, \widehat{\theta}).$$

One can interpret this quantity as follows: what is the risk of the most optimal estimator ($\inf_{\widehat{\theta}}$ part) that can minimize the risk in the worst case($\sup_{\theta}$ part)? This risk can be regarded as a *lower bound* of the performance for this risk functional minimization problem, *i.e.*, no estimator can do better than this quantity. Particularly, if for some $\alpha > 0$

$$n^{-\alpha} \asymp \inf_{\widehat{\theta}} \sup_{\theta} R(\theta, \widehat{\theta}),$$

then $n^{-\alpha}$ is called a minimax optimal rate.

If there exists an estimator $\widetilde{\theta}$ which achieves the same rate of n with the minimax risk, *i.e.*,

$$\sup_{\theta} R(\widetilde{\theta}, \theta) \asymp n^{-\alpha} \asymp \inf_{\widehat{\theta}} \sup_{\theta} R(\theta, \widehat{\theta}),$$

then such $\widetilde{\theta}$ is called a *minimax optimal estimator*. One can think this estimator as the ideal estimator at least for the sufficiently large n.

Goal One can ask the following question: in the case of generative model, for some reasonable choice of risk R, is there an algorithm that guarantees the minimax optimal estimator? Oko et al. [2023] answers this question by stating that a diffusion model, one of the most widely used generative models nowadays, achieves the nearly minimax optimal rate when risk functional is chosen by either total variation distance or Wasserstein-1 distance.

In this report, we will focus on Wasserstein-1 distance risk, and explain how a diffusion model acheieves the minimax rate.

1.2 Preliminaries

Diffusion model For some target probability measure P^*, suppose we are given only data $X_i \overset{i.i.d}{\sim} P^*$. In generative models, we want to generate the estimator of samples $\widehat{X}_i$ that has a distribution close to P^*. A diffusion model generates $\widehat{X}_i$ by the following procedures:

1. Assume P^* to be an initial distribution of some Ornstein–Ulhenbeck (OU) process, *i.e.*, for $X_0 \sim P_0 = P^*$,
$$dX_t = -\beta_t X_t dt + \sqrt{2\beta_t} dB_t.$$
Note $X_t \to N(0, I)$ exponentially. We consider this process up to some timestep T.
2. Let $Y_0 \sim N(0, I)$. The goal is to construct a dynamical system Y_t s.t. $Y_t = X_{T-t}$ so that $Y_T = X_0 = X^*$.
3. Then, it is known the following SDE induces $Y_t = X_{T-t}$ (reverse process):
$$dY_t = \beta_{T-t}(Y_t + 2\nabla \log P_{T-t}(Y_t))dt + \sqrt{2\beta_{T-t}} dB_t.$$
4. Y_T is the random variable we desire, but we cannot obtain this as we do not know the P_t. Instead, assume for each t we have $\widehat{s}(Y_t, t)$, an estimator of the score function $\nabla \log P_t(Y_t)$. Consider the estimator $\widehat{Y}_t$:
$$d\widehat{Y}_t = \beta_{T-t}(\widehat{Y}_t + 2\widehat{s}(\widehat{Y}_t, T-t))dt + \sqrt{2\beta_{T-t}} dB_t.$$
5. To generate $\widehat{X}_i \sim P^*$, set $\widehat{Y}_0^{(i)} \overset{i.i.d}{\sim} N(0, I)$. Then, $\widehat{X}_i = \widehat{Y}_T^{(i)} \approx P^*$, given $\widehat{s}$ is a *nice* estimator.

To obtain a *nice* estimator $\widehat{s}$, one find a function that minimizes the *score matching loss*, *i.e.*, for some fixed $\epsilon > 0$ and a function class $\mathcal{S}$,

$$\begin{aligned} \widehat{s} &= \operatorname*{argmin}_{s \in \mathcal{S}} \frac{1}{n} \sum_{i=1}^{n} \int_{\epsilon}^{T} \mathbb{E}_{x_t \sim P_t(x_t | (x_{0,i})_{i=1,\ldots,n})} \left[\| s(x_t, t) - \nabla \log P_t(x_t | (x_{0,i})_{i=1,\ldots,n}) \|^2 \right] dt \\ &\approx \operatorname*{argmin}_{s \in \mathcal{S}} \mathbb{E}_{x_0 \sim P^*} \left[\int_{0}^{T} \mathbb{E}_{x_t \sim P_t(x_t | x_0)} \left[\| s(x_t, t) - \nabla \log P_t(x_t | x_0) \|^2 \right] dt \right]. \end{aligned} \tag{1}$$

Note we set the target as $\nabla \log P_t(x_t|x_0)$ instead of $\nabla \log P_t(x_t)$; this trick is sometimes called a score matching trick. The ϵ is introduced to stabilize the score loss, as the loss explodes as $\epsilon \to 0$.

Typically, $\mathcal{S}$ is set to be a deep neural network in practice. Here, for theoretical analysis we will consider $\mathcal{S}$ to be a deep neural network (DNN) function class $\Phi(L, W, S, B)$ defined as follows:

$$\Phi(L, W, S, B)(x) = \left[\left(W^{(L)}(\cdot) + b^{(L)}\right) \circ \sigma \cdots \circ \sigma \circ \left(W^{(1)}(\cdot) + b^{(1)}\right)\right](x) \tag{2}$$

where

- σ: ReLU activation function.
- L: Neural network depth.
- W: Neural network width, *i.e.*, $W^{(l)} \in \mathbb{R}^{W \times W}$, $b^{(l)} \in \mathbb{R}^{W}$ for all $l = 1, \ldots, L$.
- S: Sparsity parameter, *i.e.*, $\sum_{l=1}^{L} \left[\left\|W^{(l)}\right\|_0 + \left\|b^{(l)}\right\|_0\right] \leq S$.
- B: Norm constraint, *i.e.*, $\max_{l=1,\ldots,L} \left[\left\|W^{(l)}\right\|_\infty, \left\|b^{(l)}\right\|_\infty\right] \leq B$.

We want to note that this choice of DNN is mainly due to the theoretical tractability, and is quite different to what we use in practice.

Besov space Before we move on to the main idea of the paper, we need some preliminaries regarding the function estimation problem. This is because one can think of obtaining the estimator $\widehat{X}_i$ as obtaining the estimator of the probability density function. In classical nonparametric estimation theory, one cannot estimate the arbitrary free function, as one can construct any function that interpolates the data perfectly. This perfect interpolation is not desirable, as it overfits the data and cannot generalize the region outside the data. To this end, one usually impose a regularity restriction on the target function class. One of the widely used function class in such purpose is a Besov space.

Let $\Omega = [0, 1]^d$ be a d-dimensional unit cube. For $f \in L^p(\Omega)$, define the rth-modulus of smoothness as follows:

$$w_{r,p}(f, t) := \sup_{\|h\| \leq t} \|\Delta_h^r(f)\|_{L^p(\Omega)},$$

where $\Delta_h^r(f)(x) := \sum_{j=0}^{r} \binom{r}{j}(-1)^{r-j} f(x + jh)$ if $x, x + h \in \Omega$, and 0 otherwise. For example, $\Delta_h^1(f)(x) = f(x + h) - f(x)$ and $\Delta_h^2(f)(x) = f(x + 2h) - 2f(x + h) + f(x)$. Let $r = \lfloor s \rfloor + 1$, and define a Besov semi-norm

$$|f|_{B^s_{p,q}(\Omega)} := \begin{cases} \left[\int_\Omega \left(\frac{w_{r,p}(f,t)}{t^s}\right)^q \frac{dt}{t}\right]^{\frac{1}{q}} & 0 < q < \infty, \\ \sup_{t>0} \frac{w_{r,p}(f,t)}{t^s} & q = \infty. \end{cases}$$

If f satisfies $\|f\|_{B^s_{p,q}(\Omega)} := \|f\|_{L^p(\Omega)} + |f|_{B^s_{p,q}(\Omega)} < \infty$, then f is said to be in a Besov space $B^s_{p,q}(\Omega)$. One can understand s behaves like a smoothness parameter, as in s-Sobolev space, and p behaves like a integral parameter as in L^p space.

For q, it is less direct in this definition what is the role of q is. For the interpretation of q, an equivalent definition using a wavelet decomposition may help. A wavelet theory gives the following basis decomposition for the $L^p(\Omega)$ function f:

$$f = \sum_{k \in \mathbb{Z}} \langle \phi_k, f \rangle \phi_k + \sum_{j=0}^{\infty} \sum_{k \in \mathbb{Z}} \langle \psi_{jk}, f \rangle \psi_{jk}, \text{ in } L^p, 1 \leq p \leq \infty$$

where ϕ_k denotes the father wavelet and ψ_{jk} denotes the mother wavelet at jth level. For more information about the wavelet theory, I refer Giné and Nickl [2015][Chapter 4.2].

It is known that $f \in B^s_{p,q}(\Omega)$ is equivalent to the wavelet sequence norm of f

$$\|f\|_{B^{s,W}_{pq}} \equiv \begin{cases} \|\{\langle f, \phi_k \rangle\}_k\|_p + \left(\sum\limits_{j=0}^{\infty} 2^{qj(s+\frac{1}{2}-\frac{1}{p})} \|\{\langle f, \psi_{jk} \rangle\}_k\|_p^q\right)^{1/q} & 1 \leq q < \infty \\ \|\{\langle f, \phi_k \rangle\}_k\|_p + \sup\limits_{j>0} 2^{j(s+\frac{1}{2}-\frac{1}{p})} \|\{\langle f, \psi_{jk} \rangle\}_k\|_p & q = \infty \end{cases}$$

being finite. Here, the interpretation of q is much clear. Note that $\left\|\{\langle f, \psi_{jk} \rangle\}_k\right\|$ can be interpreted as a variation at j level, which intuitively means variation at very small local domain (precisely,

2^{-j} length cube). Note this term is controlled in total by q norm. Hence, q controls some kind of aggregated amount of local variations of the function.

In sum, one can summarize the intuitive interpretation of Besov spaces as follows: $B^s_{p,q}(\Omega)$ works similar to Sobolev space $W^{s,p}(\Omega)$. q just gives a little freedom to $W^{s,p}(\Omega)$, by allowing to finely control the aggregated local variations. One can check as q gets larger the local variation can be controlled in more loose way. Particulary, $B^s_{p,1}(\Omega) \hookrightarrow W^s_p(\Omega) \hookrightarrow B^s_{p,\infty}(\Omega)$ and $B^s_{2,2}(\Omega) = W^s_2(\Omega)$.

2 Main result: Minimax optimality of the diffusion model

Now, we are ready to introduce the main result. We will introduce two results: 1. the minimax optimal rate of estimating a probability measure, and 2. the estimation error of the diffusion model generated estimator. To this end, we define $\mathcal{P}$ to be a set of absolutely continuous probability measures on Ω with a density $f \in B^s_{p,q}(\Omega)$ and $1/C \leq f \leq C$ for some $C > 0$.

2.1 Minimax optimal rate in Wasserstein distance

We first introduce the minimax optimal rate of probability measure estimator under the expected W_1 distance risk $R(P, \widehat{P}) := \mathbb{E}_{D_n} W_1(P, \widehat{P})$.

Theorem 2.1 (Minimax optimality [Niles-Weed and Berthet, 2022])**.** *For any estimator $\widehat{P} \in \mathcal{P}$ constructed using n data $D_n = (x_i)_{i=1,\ldots,n}$,*

$$n^{-\frac{s+1}{2s+d}} \lesssim \inf_{\widehat{P} \in \mathcal{P}} \sup_{P^* \in \mathcal{P}} \mathbb{E}_{D_n} W_1(P^*, \widehat{P}).$$

This theorem gives the lower bound of the minimax risk. Since there exists an estimator that achieves this precise minimax rate [Niles-Weed and Berthet, 2022][Theorem 1], this lower bound is indeed tight.

We provide a sketch of the proof:

Proof. This result is from Niles-Weed and Berthet [2022][Theorem 3, Proposition 3]. We decompose the proof into the follow steps:

Step 1. A key idea is to observe the following calculation [Maury et al., 2010]: for any $h \in C^1(\Omega)$,

$$\begin{aligned}\int_\Omega h(dP - dQ) &= \int_0^1 \frac{d}{dt}\left(\int h dP_t\right) dt = \int_0^1 \int_\Omega \nabla h \cdot v_t dP_t dt \\ &\leq \left(\int_0^1 \int_\Omega \|\nabla h\|^p dP_t dt\right)^{1/p} \left(\int_0^1 \int_\Omega \|v_t\|^q dP_t dt\right)^{1/q} \\ &\leq C^{1/p} \|\nabla h\|_{L^p(\Omega)} W_q(P, Q).\end{aligned}$$

Here v_t is a vector field induced from the continuity equation $\partial_t P_t = -div(P_t v_t)$ and C is a density upper bound. For the second equality one used the integration by parts, and for the last inequality one used a dynamical formulation of the optimal transport.

Step 2. If $P, Q \in \mathcal{P}$, one can choose the optimal h to make the bound desirable. Particularly, let $h = \sum_k \kappa_k \phi_k + \sum_{j,k} \lambda_{jk} \psi_{jk}$. Then, if $\|\kappa\|_{\ell^p} \leq 1$ and $\|\lambda\|_{\ell^p} \leq 2^{-j+dj(1/2-1/p)}$, then $\|\nabla h\|_{L^p} \lesssim 1$ [Niles-Weed and Berthet, 2022][Lemma 7].

Step 3. Plug-in $P = P^*, Q = \widehat{P}$. Since $\widehat{P}, P^* \in \mathcal{P}$, one can write $dP^*, d\widehat{P}$ as a wavelet expansion, *i.e.*, $dP^* = \sum_k \alpha^*_k \phi_k + \sum_{jk} \beta^*_{jk} \psi_{jk}$ and $d\widehat{P} = \sum_k \widehat{\alpha}_k + \sum_{j,k} \widehat{\beta}_{jk} \psi_{jk}$. Plug-in these values and h to the LHS of Step 1. Up to here, one obtains

$$C^{-1/p}(\|\alpha^* - \widehat{\alpha}\|_{\ell^p} + \sup_j 2^{-j+dj(1/2-1/p)} \left\|\beta^*_j - \widehat{\beta}^*_j\right\|_{\ell^p}) \lesssim W_q(P^*, \widehat{P}),$$

which is Niles-Weed and Berthet [2022][Proposition 3].

Step 4. As in classical minimax estimation theory in Besov space, apply Assouad's Lemma for $\mathbb{E}W_1(P^*, \widehat{P})$, *i.e.*, decompose the minimax risk into the properly chosen well-separated finite coverings of $B^s_{p,q}(\Omega)$. Then, use Step 3's result to obtain the lower bound of W_1 distance between the center of coverings, which turns out to be written in terms of Hamming distance. Then, one can optimize the bound by choosing the optimal number of levels J, which is $J \asymp n^{1/(2s+d)}$. This choice of J will give the claimed bound on n. See [Giné and Nickl, 2015][Section 6.3.1] and Kerkyacharian and Picard [1992] for the general procedure and [Niles-Weed and Berthet, 2022][Section 4.3] for the precise quantities. □

The key ingredients of the proof are three: a dynamic formulation of Wasserstein distance (Step 1), wavelet representations of the density (Step 2, 3), and Besov space minimax estimation theory (Step 4).

The next theorem shows the estimation error of the estimator generated by the diffusion model.

Theorem 2.2 (Diffusion models are nearly minimax optimal [Oko et al., 2023])**.** *For any fixed $\delta > 0$, if we train the diffusion model with scores being $\Phi(L, W, S, B)$ for some L, W, S, B that depends on n, d, p, s, T, then*

$$\sup_{X_0 \sim P^* \in \mathcal{P}} \mathbb{E}_{D_n} W_1(X_0, \widehat{Y}_{T-T_0}) \lesssim n^{-\frac{s+1-\delta}{2s+d}}$$

whenever $T_0 = n^{-\frac{2(s+1)}{2s+d}}$ and $T \geq \frac{(s+1)\log n}{\min_t \beta_t (2s+d)}$.

This theorem together with the minimax bound in Theorem 2.1 indicates that a DNN based diffusion model is 'nearly' (the gap is $n^{\frac{\delta}{2s+d}}$) minimax optimal. The key thing to take into account here is the plural scores. This indicates we will use a multiple scores, and this in fact gives some improvements; a precise will be discussed later. The T_0 can be regarded similar to 0; it is to avoid the unstable behavior of the score matching loss near the 0.

Before going into the proof, we want to first make a remark about the key idea of the proof, which utilizes the structure of the diffusion model very wisely. Heuristically speaking, the key argument is as follows: the process X_t evolves from the target distribution to Gaussian white noise. The score function needs to approximate the score of such target. When t is small, the target would be Besov function spaces, and we may not expect a better performance than the minimax rate. However, as t gets larger, our target becomes standard Gaussian, which is very smooth. Therefore, one may expect the improvement when t is sufficiently large. When one uses total variation distance, Besov function estimation error dominates the entire error, but when one uses Wasserstein-1 distance, the error is averaged, and some improvements can be made. We illustrate this more rigorously after the proof.

Again, we provide a sketch of the proof:

Proof. **Step 1.** A key idea is to consider the following stochastic process: for given $s, r \in [0, T]$, let $\overline{Y}^s(r)_t$ be a stochastic process s.t. $\overline{Y}^s(r)_0 = P_r$ and

$$d\overline{Y}^s(r)_t = \begin{cases} \beta_{T-t}(\overline{Y}^s(r)_t + 2\nabla \log P_{T-t}(\overline{Y}^s(r)_t))dt + \sqrt{2\beta_{T-t}}dB_t & t \in [0, T-s], \\ \beta_{T-t}(\overline{Y}^s(r)_t + 2\widehat{s}_t(\overline{Y}^s(r)_t, T-t))dt + \sqrt{2\beta_{T-t}}dB_t & t \in [T-s, T]. \end{cases}$$

i.e., uses the true score up to $T - s$ and then use the estimated score from $T - s$. Particularly, one can think of $\overline{Y}^0(r)_t, \overline{Y}^T(r)_t$ similar to $Y_{T-r+t}, \widehat{Y}_{T-r+t}$.

The reason we consider such stochastic process is because one can bridge the difference between true process Y_t and $\widehat{Y}$; this process resembles Y_t up to $T - s$ and resembles $\widehat{Y}_t$ from $T - s$.

Using this process, one can write the estimation error as follows:

$$\mathbb{E}W_1(X_0, \widehat{Y}_{T-T_0}) \leq \underbrace{\mathbb{E}W_1(X_0, X_{T_0})}_{:=(i)} + \underbrace{\mathbb{E}W_1(\overline{Y}^{T-T_0}(T-T_0)_{T-T_0}, \widehat{Y}_{T-T_0})}_{:=(ii)} + \underbrace{\mathbb{E}W_1(X_{T_0}, \overline{Y}^{T-T_0}(T-T_0)_{T-T_0})}_{:=(iii)}.$$

We will bound the each term separately.

Step 2. We first show $(i) \lesssim \sqrt{T_0}$. Recall by the property of OU process, one has $X_t | X_0 \sim N(m_t X_0, \sigma_t^2)$, where $m_t = \exp\left(-\int_0^t \beta_s ds\right)$ and $\sigma_t^2 = 1 - \exp\left(-2\int_0^s \beta_s ds\right)$. Particularly, $1 -$

$m_t \asymp \min\{1, t\}$ and $\sigma_t \asymp \min\left\{1, \sqrt{t}\right\}$. Then, for $Z \sim N(0, I)$ observe the following:

$$\begin{aligned}(i) &= \mathbb{E}_{X_0, X_{T_0}} \|X_0 - X_{T_0}\| \leq \mathbb{E}_{X_0}\mathbb{E}\left[\|X_0 - X_{T_0}\| \,|X_0\right] = \mathbb{E}\|X_0 - m_{T_0}X_0 + \sigma_{T_0}Z\| \\ &\leq (1 - m_{T_0})\mathbb{E}\|X_0\| + \sigma_{T_0}\mathbb{E}\|Z\| \leq (1 - m_{T_0})\sqrt{d} + \sigma_{T_0}\sqrt{d} \lesssim \sqrt{T_0}\end{aligned}$$

given $T_0 \asymp n^{-O(1)}$.

Step 3. We next bound $(ii) \lesssim \exp\left(-\min_t \beta_t T\right)$. To this end, observe $\widehat{Y}_{T-T_0}$ and $\overline{Y}^{T-T_0}(T - T_0)_{T-T_0}$ only differs in the initial distribution, $N(0, I)$ and P_T respectively. Using the fact that the OU process converges to $N(0, I)$ exponentially with respect to t in KL-divergence, and from the setting that Ω is an unit cube, one obtains

$$(ii) \lesssim TV(N(0, I), P_T) \leq \sqrt{2D_{KL}\left(P_T \,\|\, N(0, I)\right)} \lesssim \exp\left(-\min_t \beta_t T\right).$$

Step 4. We lastly bound (iii), which is the most complicated part.

Step 4-1. Consider partitioning $[T_0, T - T_0]$ by $t_j = c^j n^{-\frac{2-\delta}{2s+d}}$ for some $j = 1, \ldots, k = O(\log n)$. c is a constant chosen to satisfy $t_k = T - T_0$ for the given k. Note that the partition is *not* equipartitioned. For smaller t the interval is smaller. This choice of the partition enables us to control the error in small t region.

Step 4-2. One can decompose C as follows:

$$(iii) \leq \sum_{j=1}^{k} \mathbb{E}W_1\left(\overline{Y}^{t_{j-1}}(T - T_0)_{T-T_0}, \overline{Y}^{t_j}(T - T_0)_{T-T_0}\right).$$

Observe $\overline{Y}^{t_{j-1}}(T - T_0)_{T-T_0}$ and $\overline{Y}^{t_j}(T - T_0)_{T-T_0}$ have the same initial distribution as well as the dynamics, except the difference in the drift term in $[t_{j-1}, t_j]$. Using this property, one can derive the following bound:

$$\mathbb{E}_{X_0}W_1\left(\overline{Y}^{t_{j-1}}(T - T_0)_{T-T_0}, \overline{Y}^{t_j}(T - T_0)_{T-T_0}\right) \lesssim \sqrt{t_j \log n \int_{t_{j-1}}^{t_j} \mathbb{E}_{X_0}\left\|\widehat{s}_j(X_t, t) - \nabla \log P_t(X_t)\right\|^2 dt} + n^{-\frac{s+1}{2s+d}}. \tag{3}$$

Since the derivation of this bound is very technical, I will just explain the heuristic idea about the bound. Note the Wasserstein-1 distance is a averaged transport distance. Therefore, it can be characterized as 'Total Mass to Transport × Total Distance Transported'. Using the fact we are in the bounded domain Ω, one can show 'Total Mass to Transport' is bounded the half of the Total variation distance between two processes, which is bounded by $\sqrt{\int_{t_{j-1}}^{t_j} \cdot dt}$ term. On the other hand, using the fact that these two processes only differs in the drift term, one can show 'Total Distance Transported' is $O(\sqrt{t_j}\log n)$ with the probability at least $1 - n^{-O(1)}$. The final additive term comes from considering the probability $n^{-O(1)}$ part with total variation bound 1.

This bound represents that, as $t_j \to 0$, while score matching gets more difficult (due to the lack of regularity compared to Gaussian), its contribution to the Wasserstein-1 error is reduced. Also, this bound enables us to convert the W_1 bound into the score function estimation error bound.

Step 4-3. The estimation error of the trained score $\widehat{s}_j \in \Phi(L, W, S, B)$ at $[t_{j-1}, t_j]$ has the following bound:

$$\int_{t_{j-1}}^{t_j} \mathbb{E}_{X_0}\left\|\widehat{s}_j(X_t, t) - \nabla \log P_t(X_t)\right\|^2 dt \lesssim \left(n^{-\frac{2(s+1)}{2s+d}} \log n + t_j^{-\frac{d}{2}} n^{\frac{(\delta-2)d-4s}{2(2s+d)}} (\log n)^8\right). \tag{4}$$

This comes from the variant of the Besov function estimation theory. The derivation of this bound is deferred to Theorem 2.4.

Step 4-4. Plug-in Equation (4) to Equation (3), and recall $k \asymp \log n$. Then some algebraic calculations lead to

$$(iii) \lesssim n^{-\frac{s+1-\delta}{2s+d}}.$$

Step 5. In sum, we obtain

$$\mathbb{E}W_1(X_0, \widehat{Y}_{T-T_0}) \lesssim \sqrt{T_0} + \exp\left(-\min_t \beta_t T\right) + n^{-\frac{s+1-\delta}{2s+d}}.$$

Then, the chosen T_0, T induce the desired bound. □

Remark 2.3.

1. *When one uses the total variation distance instead of Wasserstein-1 distance, the $\sqrt{t_j}$ term does not appear in Equation (3). Therefore, as in W_1 case, the score matching gets more difficult as time becomes the small, but unlike W_1 case, its contribution cannot be controlled by partition size. Therefore, the score estimation error directly transferred to the TV distance rate, and one achieves the nearly minimax rate of Besov function estimation problem [Oko et al., 2023][Theorem 5.1]. This is worse than W_1 distance. The reason TV is worse than W_1 is that TV is a supremum based distance while W_1 is average based distance. $\sqrt{t_j \log n}$ term appears when taking an average by total distances moved in W_1. One can think this as 'Mean $\leq$ Max' type inequality.*
2. *The term δ can be any $\delta > 0$, but it cannot be 0 to obtain the bound in Equation (3). While the smaller δ makes the bound tighter, smaller δ implies one needs a finer partition, and therefore more score estimators; see t_j term has a dependency on δ. Therefore, there is a trade-off between the tighter bound and computational costs.*

We again summarize the key idea of the proof. The diffusion process structure becomes favorable if one can convert the error between distributions to the differences between either initial distributions or drift terms. Therefore, by introducing the suitable random process that can bridge the target process and the estimated process, one can convert the error by initial distributions or drift terms, and from here known theories are applicable in a nice manner.

2.2 Estimation error bound of the score estimator

Lastly, we derive the estimation error bound of the score estimator, stated in Equation (4).

Theorem 2.4 (Score estimator error [Oko et al., 2023]). *For suitably chosen L, W, S, B which depends on n, d, s, δ, t_j, the score estimator $\widehat{s}_j \in \Phi(L, W, S, B)$ defined by Equation (1) at $[t_{j-1}, t_j]$ satisfies*

$$\int_{t_{j-1}}^{t_j} \mathbb{E}_{X_0} \left\|\widehat{s}_j(X_t, t) - \nabla \log P_t(X_t)\right\|^2 dt \lesssim \left(n^{-\frac{2(s+1)}{2s+d}} + t_j^{-\frac{d}{2}} n^{\frac{(\delta-1)d-2s}{2s+d}} (\log n)^8\right).$$

Before moving on to actual proof, we introduce the general strategy to obtain the function estimation error bound when one has the estimator $\widehat{f}$ for the target f. The general procedure decomposes into two: 1. Approximation stage, and 2. Statistical learning stage.

In approximation stage, one tries to obtain the ideal approximation error between two function classes, *i.e.*,

$$\sup_{f \in B^s_{p,q}(\Omega)} \inf_{\widetilde{f} \in \Phi} \left\|\widetilde{f} - f\right\|^2_{L^2(P_x)}.$$

One can interpret this error as follows: since one takes the infimum first, given a target f, one tries to find the best approximator in Φ that approximates the target well. Then, by taking the supremum, one look at the worst choice of the target. In sum, one seeks for the worst error that the ideal approximator has. Note this stage does not involve any stochasticity. In this stage, Banach geometry theory will be the main tool to obtain the bound. Particularly, one typically write f as some N-term basis expansion with suitable bases. Then, we next check how an estimator class can approximate such basis expansion. In a simple form, one typically tackles approximation problem by showing 'target function class $\approx$ N-term basis expansion $\approx$ estimator class'.

In statistical learning stage, heuristically speaking, one decomposes the total error as follows:

$$\sup_f \mathbb{E}_{D_n} \left\|\widehat{f} - f\right\| \leq \sup_f \left\|f - \widetilde{f}\right\| + \sup_f \mathbb{E}_{D_n} \left\|\widehat{f} - \widetilde{f}\right\|.$$

The first term corresponds to the approximation error. Therefore, one aims to control the second term in statistical learning part. Since $\left\|\widehat{f}-\widetilde{f}\right\|$ is a stochastic process w.r.t. n, empirical process theory is used to control the supremum of this process. In practice, one may use squared norm rather than plain norm which does not have a triangular inequality, but similar decomposition of the bound is used.

Once one obtains both approximation and statistical learning, in the total error one optimizes N, a number of basis, by the n, a number of data. This gives the total estimation error in terms of n.

Keeping this overall strategy in mind, we will proceed with the sketch of the proof.

Proof. **Step 1.** We obtain the approximation result here. We will briefly go over the following statement [Oko et al., 2023][Lemma 3.6]:

Let $N \gg 1$ and $t_{j-1} \geq N^{-(2-\delta)/d}/2$. Then, there exists $\widetilde{s} \in \Phi(L, W, S, B)$ for some L, W, S, B that depends on N such that for all $t \in [t_{j-1}, t_j]$

$$\mathbb{E}_{X_t} \left\|\widetilde{s}(X_t, t) - \nabla \log P_t(X_t)\right\|^2 \lesssim \frac{1}{\sigma_t^2} N^{-\frac{2(s+1)}{d}}.$$

Here, N can be interpreted as taking a N-terms basis expansion of the Besov function. Instead of rigorous proof, we just heuristically explain the key idea of the proof here.

First, before go into the score function approximation problem, assume we are only in a plain Besov function estimation problem. In this case, so-called B-spline functions turned out to be a suitable choice of basis [Devore and Popov, 1988], *i.e.*, $B^s_{p,q} \approx$ B-spline. To construct a B-spline function, one consider the following m order piecewise polynomial:

$$N_m(x) = \left(\underbrace{\mathbb{1}_{[0,1]} * \cdots * \mathbb{1}_{[0,1]}}_{(m+1) \text{ times}} \right)(x).$$

Then, a B-spline function is written as follows:

$$M^{m,d}_{k,j}(x) := \prod_{i=1}^{d} N_m(2^{k_i} x_i - j_i).$$

Suzuki [2019] showed this function can be approximated well by $\Phi(L, W, S, B)$. The idea comes from Yarotsky [2017], which showed a product function $\prod_i x_i$ can be approximated well by Φ. In fact, one can write N_m as a linear combination of product functions, and $M^{m,d}_{k,j}$ is a product function of N_m. Therefore, by applying product function approximation repeatedly, the approximation $\Phi(L, W, S, B) \approx$ B-spline can be shown. Here, the choice of L, W, S, B depends on N, d, s.

Aggregating two approximation results $B^s_{p,q} \approx$ B-spline $\approx \Phi$, the rate w.r.t. N can be derived for the plain Besov functions; the precise rate in this case is $N^{-\frac{2s}{d}}$.

Now, we move to the score approximation problem. At glance, if $f \in B^s_{p,q}(\Omega)$, one may expect $\nabla \log f$ to have $d+1$ dimension (includeing the time domain) and $s-1$ smoothness; it seems like we might get a weaker bound. However, here, the unique structure of the diffusion model, the combination of OU process and score loss, turned out to be effective. To observe this, first notice due to the OU process structure, $P(X_t|X_0) \sim N(m_t X_0, \sigma_t^2)$ for some m_t and σ_t (which was in fact introduced in the above). This enables us to write $P_t(x) = \int P_0(y) K_{\sigma_t^2}(\|x - m_t y\|^2) dy$ where K is a Gaussian kernel. Then, the score $\nabla \log P_t(x)$ can be also written as a fraction of $B^s_{p,q} * K_{\sigma_t}$.

From this observation, one can naturally extend the plain Besov function approximation method to this setting: substitute the role of $M^{m,d}_{k,j}$ in the plain Besov approximation to

$$E^{m,d}_{j,k}(x, t) = \prod_{i=1}^{d} \int N_m(2^{k_i} x_i - j_i) P_{N(m_t y_i, \sigma_t^2)}(x_i) dy_i.$$

This looks very natural extension of plain Besov function approximation to this score function approximation. This intuition turns out to be true. From here, one can proceed with the similar technique in Besov function approximation problem to derive the desired bound. Again, to control the approximation error we have the precise quantity of L, W, S, B with respect to N, d, s.

Step 2. Now, we obtain the similar statement to the total error decomposition as discussed in the above. Particulary, Oko et al. [2023][Theorem C.4] gives the desired result: if $\widehat{s}$ is an score estimator defined by an empirical risk minimizer, then for all $\eta > 0$

$$\mathbb{E}_{D_n\sim P_0}\int_{t_{j-1}}^{t_j}\mathbb{E}_{X_t\sim P_t}\left\|\widehat{s}(X_t,t)-\nabla\log P_t(X_t)\right\|^2 dt \le$$

$$2\inf_{\widetilde{s}\in\Phi}\int_{t_{j-1}}^{t_j}\mathbb{E}_{X_t\sim P_t}\left\|\widetilde{s}(X_t,t)-\nabla\log P_t(X_t)\right\|^2 dt+\frac{2c}{n}\left(\frac{37}{9}\log N(\mathcal{L},\|\cdot\|_{L^\infty(\Omega)},\eta)+32\right)+\eta.$$

Here, $\mathcal{L}=\left\{\int_{t_{j-1}}^{t_j}\mathbb{E}_{X_t}\left\|s(X_t,t)-\nabla\log P_t(X_t|X_0)\right\|^2 dt \mid s\in\Phi(L,W,S,B)\right\}$ is a function space for the loss function, and $N(\mathcal{L},\|\cdot\|_{L^\infty},\eta)$ is a covering number. This bound can be derived using a similar technique in Schmidt-Hieber [2020][Lemma 4]. Since this part is not a main contribution of this paper, we omit the explanation of this bound.

Step 3. Plug-in the approximation rate $N^{-\frac{2(s+1)}{d}}$, $\eta = n^{-\frac{2(s+1)}{2s+d}}$ and the covering number bound of $\log N(\mathcal{L},\|\cdot\|_{L^\infty},\eta) \lesssim SL\log\left(\frac{LWBn}{\eta}\right)$ which is derived in Oko et al. [2023][Lemma C.2] whenever $\|s\|_{L^\infty}\lesssim\log n$. This covering number bound is derived from the fact that ReLU activation was used, so that one can apply Lipschitz continuity repeatedly. Since L, W, S, B is determined in terms of N from Step 1, we can write this total error in terms of n and N. Lastly, one optimizes the choice of N with respect to n, yielding the desired bound. □

Remark 2.5. *In Step 1, the approximation stage, we observed the score approximation problem can be written as approximating the function of the form $B^s_{p,q} * K_{\sigma_t}$. one can think of this approximation being almost equivalent to approximating $B^s_{p,q}$ and then take a smooth Gaussian convolution. Therefore, one does not have to deal with $d+1$ dimension and $s-1$ smoothness, but just deal with d dimensional s smoothness problem; in fact, by taking a convolution with Gaussian kernel we gain additional smoothness, and this is revealed in the bound: score approximation bound $N^{-\frac{2(s+1)}{d}}$ turned out to be faster than $N^{-\frac{2s}{d}}$.*

The novelty of the proof comes from considering some modification to the B-spline basis; the authors refer it as *diffused B-spline basis*. This modification to B-spline basis was doable due to the Gaussian conditional distribution formulation which comes from OU process structure.

3 Strength and weakness of the paper

I would like to summarize the key strength of Oko et al. [2023] as follows:

1. Use of unique structure of diffusion models: The diffusion model differs from other types of generative models (e.g., GAN, VAE) in the sense that diffusion model ‘incrementally’ add/remove noises; other types of generative models rather directly transform the noise. Diffusion models turned out to perform much better than other models, while it was not clear why this specific structure of diffusion model helps. In this proof, one can find the advantage of this incremental noise addition/removal: Equation (3) shows by incrementally adding the noise, by choosing the score differently for each partition, one can control the ‘non-smooth’ target function error part small. This gives one possible explanation why incrementally using the noise is better than directly transforming the noise.
2. Novel bases approximation for score approximation: Authors used the favorable property of the score function that it can be written as a convolution between Besov function and Gaussian kernel, and constructed unique basis that is perhaps most suitable to the score function class. This modified basis is tailor-made to OU process structure and gives additional advantage on the rate. The fact DNNs can approximate this tighter basis can also be another possible answer why diffusion models outperform other generative models.
3. Use of OU process structure: The technique of introducing some bridge stochastic processes and bound the distance using either deviation from the initial distribution or drift term was

new to me. This seems like very nice idea to apply in general when one wants to deal with more general diffusion processes.

That said, I found some limitations of the paper:

1. The structure of $\Phi(L, W, S, B)$: Their theory crucially relies on the structure of $\Phi(L, W, S, B)$ used in Yarotsky [2017]. However, This function class hard to use in practice, as one have to solve some constraint optimization, particularly with S and B. This constraint works in two parts. In approximation stage this constraint avoids some overfitting to the noise. In statistical learning stage, this constraint enables covering number bound to hold. I believe this kind of constraint can be avoided by adding some kind of explicit regularizer or thresholding methods, which have been studied a lot in nonparametric statistics. These alternative approaches are more desirable in practice as these are easier to implement.
2. Adaptivity: The optimal choice of L, W, S, B, T in this paper relies on s, the smoothness parameter of the target distribution. In practice this quantity is unknown. In this regard, the DNN estimator is not adaptive; we may not expect such performance when we do not know the smoothness parameter s. Bayesian methods or boosting type methods can be one alternative to think of to overcome this problem.

4 Discussions

Overall, the paper nicely deals with the theoretical property of diffusion model estimators. In particular, their proof gives one potential answer why diffusion models outperform other types of generative models: fine partition and Gaussian convolution formulations enable DNNs to utilize the additional smoothness of the Gaussian noise to obtain the minimax optimal bound.

On the other hand, in a practical point of view, their estimator may not be tractable; it is unclear diffusion models used in practice work in a same manner, as they are different estimators to the proposed estimator in the paper. In addition, their estimator is not adaptive, opening the question whether one can obtain the adaptive estimator in this generative model setting.

References

R. A. Devore and V. Popov. Interpolation of besov spaces. *American Mathematical Society*, 305:397 – 414, 1988.

David L. Donoho and Iain M. Johnstone. Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3):879 – 921, 1998. doi: 10.1214/aos/1024691081. URL https://doi.org/10.1214/aos/1024691081.

Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2015.

G. Kerkyacharian and D. Picard. Density estimation in besov spaces. *Statistics Probability Letters*, 13(1):15–24, 1992. ISSN 0167-7152. doi: https://doi.org/10.1016/0167-7152(92)90231-S. URL https://www.sciencedirect.com/science/article/pii/016771529290231S.

Bertrand Maury, Aude Roudneff-Chupin, and Filippo Santambrogio. A macroscopic crowd motion model of gradient flow type, 2010. URL https://arxiv.org/abs/1002.0686.

Jonathan Niles-Weed and Quentin Berthet. Minimax estimation of smooth densities in Wasserstein distance. *The Annals of Statistics*, 50(3):1519 – 1540, 2022. doi: 10.1214/21-AOS2161. URL https://doi.org/10.1214/21-AOS2161.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023. URL https://openreview.net/forum?id=6961CeTSFA.

Anselm Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *Annals of statistics*, 48(4):1875–1897, August 2020. ISSN 0090-5364. doi: 10.1214/19-AOS1875.

Taiji Suzuki. Adaptivity of deep reLU network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1ebTsActm.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks, 2017. URL https://arxiv.org/abs/1610.01145.