Minimax optimality of diffusion models

Jiyoung Park

October 10, 2025

TEXAS A&M UNIVERSITY

- Let P^* be an arbitrary target measure. Set $X_0 \sim P^*$ and construct a diffusion model from n data D_n. Let $\widehat{Y}_T$ be the random element induced from the diffusion model after the sufficient iterations T. Then, what would be the worst case estimation rate, *i.e.*,

$$\sup_{X_0 \sim P^*} \mathbb{E}_{D_n} d\left(X_0, \widehat{Y}_T\right) \lesssim n^{-\square}?$$

- How optimal the above rate is?
    - For $\widehat{P}$ any estimator of P^*, the following rate is called 'minimax optimal rate':

$$\inf_{\widehat{P}} \sup_{P^*} \mathbb{E}_{D_n} d(P^*, \widehat{P}) \gtrsim n^{-\square}.$$

    - Can diffusion model achieve the minimax optimal rate? YES, at least nearly, when $d = TV$ or W_1.
- I will focus on $d = W_1$; this one has more interesting intuition than TV.

Preliminaries

- Besov space: characterizes a function space with some 'smoothness'.
  - Fix the domain of X by $\Omega := [0,1]^d$, a unit hypercube.
  - For $f \in L^p(\Omega)$, define the 'rth-modulus of smoothness':

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

  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.
    - E.g. $\Delta_h^1(f)(x) = f(x+h) - f(x)$; $\Delta_h^2(f)(x) = f(x+2h) - 2f(x+h) + f(x)$.
  - 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 $B^s_{p,q}(\Omega)$.
- Besov space is not a Banach space, but quasi-Banach space.
- Easier interpretaions by examples:
  - $B^s_{p,1}(\Omega) \hookrightarrow W^s_p(\Omega) \hookrightarrow B^s_{p,\infty}(\Omega)$. Particulary, $B^s_{2,2}(\Omega) = W^s_2(\Omega)$.
  - $B^s_{p,q} \approx W^s_p$, and q is just for some finer distinctions.
  - As in Sobolev embedding, $s > d/p$ implies the continuity of f.
  - Important example for later: $B^1_{\infty,1}(\Omega) \hookrightarrow Lip(\Omega) \hookrightarrow B^1_{\infty,\infty}$.

- Let $\Phi(L, W, S, B)$ be a L-layer W-width ReLU Deep neural network with the following structure:

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

  - 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 $l = 1, \dots, L$.
  - 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.*, $\max_{l=1,\dots,L} \left[\left\| W^{(l)} \right\|_\infty, \left\| b^{(l)} \right\|_\infty \right] \leq B$.

- Goal: Given only data $x_i \overset{i.i.d}{\sim} P^*$, generate more samples $x_i \sim P^*$.
- Procedure:
  1. Assume P^*: 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. $Y_t = X_{T-t} \Rightarrow Y_T = X_0 = X^*$.
  3. Then, 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 distribution 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 $x_i \sim P^*$, set $\widehat{Y}_0^{(i)} \overset{i.i.d}{\sim} N(0, I) \Rightarrow x_i = \widehat{Y}_T^{(i)} \approx P^*$, given $\widehat{s}$ is a *nice* estimator.

- To obtain the *nice* estimator $\widehat{s}(Y_t, T-t)$, one trains a function (typically DNN) with 'score mathcing loss'. Fix some function class $\mathcal{S}$ (typically DNN); then, for some fixed $\epsilon > 0$, define

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

$$\approx \underset{s \in \mathcal{S}}{\operatorname{argmin}} \mathbb{E}_{x_0 \sim P^*} \left[\int_{0}^{T} \mathbb{E}_{x_t \sim P_t(x_t|x_0)} \left[\| s(x_t, t) - \nabla \log P_t(x_t|x_0) \|^2 \right] dt \right].$$

- 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.

TEXAS A&M UNIVERSITY

- Let $\mathcal{P}$ to be a set of absolutely continuous probability measures on Ω with density f in a Besov space $B^s_{p,q}(\Omega)$ and bounded below and above by C_f^{-1} and C_f.

Theorem (Minimax optimality (NWB22))

For any estimator $\widehat{P} \in \mathcal{P}$ constructed using n data $D_n = (x_i)_{i=1,\dots,n}$,

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

Theorem (Diffusion models are nearly minimax optimal (OAS23))

For any $\delta > 0$, if we train the diffusion model with the score estimator $\widehat{s}(x,t) \in \Phi(L, W, S, B)$ for some L, W, S, B that depends on n, d, p, s, T and $T \geq \frac{(s+1)\log n}{\min_t \beta_t(2s+d)}$, then

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

- DNN diffusion model is 'nearly' (the gap is $n^{\frac{\delta}{2s+d}}$) minimax optimal.
- In practice, the score loss blows up as $t \to 0$, so often one uses clipping $t \in [\epsilon, T]$ for some $\epsilon > 0$. The actual theorem is written w.r.t $\widehat{Y}_{T-\epsilon}$, but for simplicity we assume $\epsilon = 0$.

Proof idea: Minimax optimality

TEXAS A&M UNIVERSITY

- Result is from (NWB22)[Theorem 3, Proposition 3].
- Key idea: Observe the following calculation (MRCS10): for any $h \in C^1(\Omega)$,

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

- If $P, Q \in \mathcal{P}$, one can choose the optimal h in LHS to get $\|f_P - f_Q\|_{B^{-1}_{1,\infty}} \lesssim W_1(P, Q)$.
- Plug-in $P = P^*, Q = \widehat{P}$, and $\left\|f_{P^*} - f_{\widehat{P}}\right\|_{B^{-1}_{1,\infty}}$'s lower bound can be derived using the standard Besov space minimax estimation technique (KP92).

TEXAS A&M UNIVERSITY

- To prove the minimax optimality of the DNN diffusion model, we first need the performance guarantee of DNN in general Besov function estimation.
- Consider the problem of estimating $f^* \in B_{B^s_{p,q}(\Omega)} \cap B_{L^\infty(\Omega)}(0, F)$ for some $F > 0$, with the data $y_i = f^*(x_i) + \epsilon_i$ with $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$ and $X \sim P$ where $supp(P) \subseteq \Omega$.

Theorem (DNN estimator of Besov function)

Let $\widehat{f} := \text{argmin}_{h \in \Phi(L,W,S,B)} \sum_{i=1}^n |y_i - h(x_i)|^2$ *with* L, W, S, B *that depends on* n, s, d, p. *For all* $f^* \in B_{B^s_{p,q}(\Omega)}(0,1) \cap B_{L^\infty(\Omega)}(0, F)$ *with some* $F > 0$*,*

$$\mathbb{E}_{D_n} \left\| f^* - \widehat{f} \right\|^2_{L^2(P)} \lesssim n^{-\frac{2s}{2s+d}} (\log n)^3.$$

- The proof consists of two ingredients:
  - Approximation of Besov function by some DNN $\widetilde{f}$ (may depends on f^*).
  - Statistical learning theory to control the error between $\widehat{f}$ and any choice of the approximator $\widetilde{f}$.
  - Total error is bounded by the above two errors $\left\| \widehat{f} - \widetilde{f} \right\|, \left\| \widetilde{f} - f^* \right\|$.

- Optimal approximation error: For sufficiently large $N \in \mathbb{N}$, there exists L, W, S, B that depends on N, d, s, p s.t.

$$\sup_{f^* \in B_{B^s_{p,q}(\Omega)}(0,1)} \inf_{\widetilde{f} \in \Phi(L,W,S,B)} \left\| \widetilde{f} - f^* \right\| \lesssim N^{-\frac{s}{d}}.$$

- Basic strategy: two-stage approximation: $B^s_{p,q}(\Omega) \approx$ B-spline functions $\approx \Phi(L, W, S, B)$.
  - B-spline functions:
    - Fix m and consider

$$N_m(x_i) := \Big(\underbrace{\mathbb{1}_{[0,1]} * \mathbb{1}_{[0,1]} * \cdots * \mathbb{1}_{[0,1]}}_{(m+1) \text{ times}} \Big)(x_i).$$

    - $N_m(x)$ is a piecewise polynomial of the order m.
    - The following basis is called B-spline.

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

One can think of j as a location parameter (like 0th Haar wavelet basis) and k as spatial resolution (like kth Haar wavelet basis).

Key ingredient: Approximation II

- $B^s_{p,q}(\Omega) \approx$ B-spline is established in (DP88).
- B-Spline $\approx \Phi(L, W, S, B)$ is from the following observations:
  - For some $M > 0$, write $\phi_{(0,M)}(x) := \sigma(x) - \sigma(x - M) = M \wedge \sigma(x)$.
  - Observe $N_m(x)$ has the form

$$N_m(x) = \frac{1}{m!} \sum_{j=0}^{m+1} (-1)^j \binom{m+1}{j} (m+1)^m \left(\phi_{(0,1-\frac{j}{m+1})} \left(\frac{x-j}{m+1} \right) \right)^m .$$

First, we focus on approximating $\left(\phi_{(0,1-\frac{j}{m+1})} \left(\frac{x-j}{m+1} \right) \right)^m$.

  - (Yar17) showed for some $D \in \mathbb{N}$ there exists $\psi : \mathbb{R}^D \to \mathbb{R} \in \Phi(L_1, W_1, S_1, B_1)$ for some L_1, W_1, S_1, B_1 that depends on m and ϵ such that

$$\sup_{x \in [0,M]} \left| \psi \underbrace{\left(\phi_{(0,M)} \left(\frac{x}{M} \right), \ldots, \phi_{(0,M)} \left(\frac{x}{M} \right) \right)}_{m \text{ times. Write this function as } \psi \circ \phi_{(0,M)}(x/M).} - \left(\phi_{(0,M)} \left(\frac{x}{M} \right) \right)^m \right| \leq \epsilon$$

  - Therefore, the reasonable construction of the approximator of $N_m(x)$ will be

$$f(x) = \frac{1}{m!} \sum_{j=0}^{m+1} (-1)^j \binom{m+1}{j} (m+1)^m \left(\psi \circ \phi_{(0,1-\frac{j}{m+1})} \left(\frac{x-j}{m+1} \right) \right) .$$

  - Then, appropriately using ψ and f makes the form of $M^{m,d}_{0,0}(x)$.

- For any $F > 0$ and any function space $\mathcal{F} \subseteq B_{L^\infty(\Omega)}(0, F)$, there exists the following generalization gap type bound:

$$\mathbb{E}_{D_n} \left\| f^* - \widehat{f} \right\|^2_{L^2(P)} \leq C \left(\underbrace{\inf_{f \in \mathcal{F}} \|f^* - f\|^2_{L^2(P)}}_{\approx \left\| f^* - \widetilde{f} \right\|^2} + \underbrace{(F^2 + \sigma^2) \frac{\log N(\mathcal{F}, \delta, \|\cdot\|_\infty)}{n} + \delta(F + \sigma)}_{\approx \mathbb{E} \left\| \widehat{f} - \widetilde{f} \right\|^2} \right).$$

Proof strategy:

1. Substitute $\widehat{f}$ to the closest δ-minimal covering of $\mathcal{F}$ and use the fact $\mathcal{F} \subseteq B_{L^\infty(\Omega)}(0, F)$ to bound the population risk by the empirical risk (Hardest part).
2. Bound the empirical risk in terms of the optimal recovery error: By using the fact that $\widehat{f}$ is ERM.

- Set $\mathcal{F} = \Phi(L, W, S, B) \cap B_{L^\infty(\Omega)}(0, F)$, and then the covering number analysis will give the following:

$$\log N\left(\Phi(L, W, S, B), \delta, \|\cdot\|_\infty\right) \leq 2SL \log\left((B \vee 1)(W + 1)\right) + S \log\left(\frac{L}{\delta}\right).$$

- Set $\delta = 1/n$, and in Step 1's RHS.

- Apply (1). the approximation result to get $\inf_{f \in \mathcal{F}} \|f^* - f\|_{L^2(P)}^2 \lesssim N^{-\frac{s}{d}}$, and (2). the covering number bound obtained in Step 2 with specific L, W, S, B in the approximation result.
- Then, optimizing the RHS w.r.t. N will induce the claimed bound with $N \asymp n^{\frac{d}{2s+d}}$.

DNN estimation to the score loss

- Since we are estimating the score (log-derivative) uniformly over the time t, there is a slight modification of the above result.
- Naively, this seems like a $d+1$ dimensional and $s-1$ smoothness function estimation problem. But, there is additional information for this problem: $P(X_t|X_0) \sim N(m_t X_0, \sigma_t^2)$ for some m_t, σ_t^2.
- $\therefore P_t(x) = \int P_0(y) K_{\sigma_t^2}(\|x - m_t y\|^2) dy$ where $K_{\sigma_t^2}$ is a Gaussian kernel. Therefore, our target $\nabla \log P_t(x)$ also written as a fraction of $B_{p,q}^s * K_{\sigma_t^2}$.
- If we substitute $N_m(2^{k_i} x_i - j_i)$ in the B-spline by

$$E_{j,k}(x_i, t) = \int \mathbb{1}_{\{0,1\}}(2^{k_i} x_i - j_i) P_{N(m_t y_i, \sigma_t^2)}(x_i) dy_i,$$

  Gaussian parts and Besov density parts separately controlled each other, and one can approximate $B_{p,q}^s * K_{\sigma_t^2}$ by $E_{j,k}(x_i, t)$. One can do the similar procedure as the above with this bases.
- $\Rightarrow$ Population Score Loss of $\widehat{s} \lesssim n^{-\frac{2s}{2s+d}} (\log n)^{16}$

- Using the Besov space estimation result, one can show

$$\sup_{P^*} \mathbb{E}_{D_n} TV(X_0, \widehat{Y}_T) \lesssim n^{-\frac{s}{2s+d}} (\log n)^8.$$

- W_1 rate $n^{-\frac{s+1-\delta}{2s+d}}$ turned out to be faster. Why?
- Key observation: Utilizing the smoothness of the Gaussian noise.
  - Note the score network $s(X_t, t)$ does not have to be uniformly same over the time.
  - Observe $s_0 \approx \nabla B^s_{p,q}$, while $s_T \approx \nabla N(0, I)$.
  - Since $N(0, I)$ is very smooth, s_T is much easier to approximate/estimate than s_0.
  - $\therefore$ After the certain timestep t', estimation error is expected to be much smaller.
  - Wrong but intuitive illustration:

$$d(\widehat{P}_{[0,T]}, P^*_{[0,T]}) \leq \underbrace{d(\widehat{P}_{[0,t']}, P^*_{[0,t']})}_{\text{non-smooth target}} + \underbrace{d(\widehat{P}_{[t',T]}, P^*_{[t',T]})}_{\text{smooth target (Gaussian score)}}.$$

  When $d = TV$, the 'non-smooth' term dominates, so cannot improve the DNN estimator rate. But when $d = W_1$, non-smooth part contributes less, so there is an improvement.

TEXAS A&M UNIVERSITY

1. 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., use 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 $Y_{T-r+t}, \widehat{Y}_{T-r+t}$.

2. Our target: $\mathbb{E}W_1(X_0, \widehat{Y}_T) \leq \mathbb{E}W_1(\overline{Y}^T(T)_T, \widehat{Y}_T) + \mathbb{E}W_1(X_0, \overline{Y}^T(T)_T)$.

3. First term: $\widehat{Y}_T$ and $\overline{Y}^T(T)_T$ only differs in the initial distributions ($N(0, I)$ and P_T resp.), leading to $\lesssim TV(N(0, I), P_T) \leq \exp(-\beta T)$ ($\because$ reverse OU process).

4. Second term:
   - Discretize $[0, T]$ by the partition made by $t_j = C^j n^{-\frac{2-\delta}{2s+d}}$ for some $j = 1, \ldots, k = O(\log n)$. Here C is a constant that makes $C^k n^{-\frac{2-\delta}{2s+d}} = T$. A certain t_j will be t' mentioned above.
   - Important: This interval is not 'equi-length'. Smaller t has the smaller interval.
   - $\mathbb{E} W_1(X_0, \overline{Y}^T(T)_T) \leq \sum_j \mathbb{E} W_1(\overline{Y}^{j-1}(T)_T, \overline{Y}^j(T)_T)$.
   - $\overline{Y}^{j-1}(T)_T$ and $\overline{Y}^j(T)_T$ has the same initial distribution as well as the dynamics, except the difference in the drift term of $[t_{j-1}, t_j]$.
   - Girsanov Theorem gives the KL bound of such processes in terms of the difference between drift terms, and (omitting the complicated steps) leads to

$$\mathbb{E}_{x_0} W_1(\overline{Y}^{j-1}(T)_T, \overline{Y}^j(T)_T) \lesssim \sqrt{t_j \log n \int_{t_{j-1}}^{t_j} \mathbb{E}_{x_t, x_0 \sim P_t, P_0} \|\widehat{s}(x_t, t) - \nabla \log P_t(x_t)\|^2 \, dt} + n^{-\frac{s+1}{2s+d}}.$$

     Note the bound gets smaller when t_j is small (corresponding to the Key Observation).
   - Plug-in the estimation error bound of the score loss (in the interval $[t_{j-1}, t_j]$), and plug-in $t_j = C^j n^{-\frac{2-\delta}{2s+d}}$ to derive the desired value.

- The extra term δ appears in W_1 from optimizing the choice of the threshold t'.
- In case of TV distance, one obtains the bound as in the above, but with t_j part substituted by $O(1)$. So, one cannot tighten the bound when t_j is small, so the error of non-smooth part equally contributes.
- One can think of this as 'Mean-diff ($\approx W_1$) $\leq$ Max-diff ($\approx$ TV)' type inequality.

- DNN diffusion model with score mathcing loss achieves almost minimax rate w.r.t. W_1 (and TV) distance.
- The fundamental ingredient is from the minimax function estimation in Besov space.
  - Approximation theory to obtain the good approximator (B-Spline in Besov case).
  - Learning theory to bound the gap between estimator and the approximator.
- The OU process structure of the diffusion model gives some advantage:
  - Target score has the same smoothness as the density.
  - As $t \to T$, target score gets smoother.

- How to actually train such a constraint neural network?
  - Can we reformulate the constraint into tractable ways, e.g., unconstraint optimization with appropriate regularizer?
  - Constraints play a role in two parts:
    1. Approximation: S, B enables to avoid the overfitting to the noise ($\approx$ LASSO type regularizer).
    2. Learning: S, B enables to bound the covering number, which controls the generalization bound.
  - It is not immediate how to avoid such constraints in the approximation stage: Relationship to weight decay type penalty?
  - On the other hand, there are alternative approaches to obtain a generalization bound (e.g., Rademacher complexity) to avoid the constraint. Can we utilize those?
  - PAC (Bayes) type analysis for the specific algorithm?
- Adaptivity
  - Constructing $\Phi(L, W, S, B)$ requires the prior knowledge on the regularity of the P^*; e.g., choices of L, W, S, B require s, p. This makes the estimation non-adaptive.
- Claims using 'two' DNNs at $[0, t']$ and $[t', T]$ improves the rate. When to exactly? Can $\widehat{s}(x_t, t)$ be adaptive to t'?

Thank You For Your Attention!

[DP88] R. A. Devore and V. Popov, *Interpolation of besov spaces*, American Mathematical Society **305** (1988), 397 – 414.

[KP92] G. Kerkyacharian and D. Picard, *Density estimation in besov spaces*, Statistics Probability Letters **13** (1992), no. 1, 15–24.

[MRCS10] Bertrand Maury, Aude Roudneff-Chupin, and Filippo Santambrogio, *A macroscopic crowd motion model of gradient flow type*, 2010.

[NWB22] Jonathan Niles-Weed and Quentin Berthet, *Minimax estimation of smooth densities in Wasserstein distance*, The Annals of Statistics **50** (2022), no. 3, 1519 – 1540.

[OAS23] Kazusato Oko, Shunta Akiyama, and Taiji Suzuki, *Diffusion models are minimax optimal distribution estimators*, ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models, 2023.

[Yar17] Dmitry Yarotsky, *Error bounds for approximations with deep relu networks*, 2017.