On Minimax Optimality of Estimating the Wasserstein Metric

Tengyuan Liang*[1]

[1]University of Chicago, Booth School of Business

August 27, 2019

Abstract

We study the minimax optimal rate for estimating the Wasserstein-1 metric between two unknown probability measures based on n i.i.d. empirical samples from them. We show that estimating the Wasserstein metric itself between probability measures, is not significantly easier than estimating the probability measures under the Wasserstein metric. We prove that the minimax optimal rates for these two problems are multiplicatively equivalent, up to a $\log\log(n)/\log(n)$ factor.

1 Introduction

In this note we study the minimax optimal rates for estimating the population Wasserstein metric between probability measures based on empirical samples. Let μ, ν be two probability measures in $\Omega = [0,1]^d$, and $W(\mu,\nu)$ denote the Wasserstein-1 distance between them. Suppose $X_1, \dots X_m$ are i.i.d samples from μ, and $Y_1, \dots, Y_n$ i.i.d from ν. We study: the minimax optimal rate for estimating $W(\mu,\nu)$ based on $\{X_i\}_{i=1}^m, \{Y_j\}_{j=1}^n$, for some class of probability measures $\mathcal{G}$ of interest

$$\inf_{\widetilde{T}_{m,n}} \sup_{\mu,\nu\in\mathcal{G}} \mathbf{E}\, |\widetilde{T}_{m,n} - W(\mu,\nu)| \ . \tag{1.1}$$

The problem is of importance in both statistics and machine learning, with applications such as nonparametric two sample testing, evaluation of the transportation cost from one set of samples to another, and transfer learning. It turns out that using empirical measures $\widehat{\mu}_m, \widehat{\nu}_n$ to estimate is a bad idea. Due to a result by Dudley (1969), even for infinitely smooth $\mathcal{G} = \{\mathrm{Unif}(\Omega)\}$ and $d \geqslant 2$,

$$\sup_{\mu,\nu\in\mathcal{G}} |W(\widehat{\mu}_m, \widehat{\nu}_n) - W(\mu,\nu)| \asymp n^{-\frac{1}{d}} \ . \tag{1.2}$$

A natural question arises: can one obtain faster rate, for estimating the Wasserstein metric with other estimators $\widetilde{T}_{m,n}$ leveraging the regularity of $\mathcal{G}$ such as smoothness.

A related yet different problem studied in the current literature is estimating a probability measure under the Wasserstein metric based on samples (Weed and Bach, 2017; Liang, 2018; Singh et al., 2018; Weed and Berthet, 2019):

$$\inf_{\widetilde{\nu}_n} \sup_{\nu\in\mathcal{G}} \mathbf{E}\, W(\widetilde{\nu}_n, \nu) \ . \tag{1.3}$$

*Liang gratefully acknowledges support from the George C. Tiao Fellowship.

The two problems are close in nature: "estimating the metric itself" is usually an **easier** problem than "estimating under the metric." In fact, the solution of the latter problem $\widetilde{\mu}_m, \widetilde{\nu}_n$ naturally induces a plug-in answer to the first, since

$$\mathbf{E} \left| W(\widetilde{\mu}_m, \widetilde{\nu}_n) - W(\mu, \nu) \right| \leqslant \mathbf{E}\, W(\widetilde{\mu}_m, \mu) + \mathbf{E}\, W(\widetilde{\nu}_n, \nu) \enspace .$$

However, it is unclear whether such a plug-in estimator is optimal. In fact, it is well-known that estimating specific functional of density $F(\nu)$ is usually strictly easier than estimating the density ν itself. For example, in estimating quadratic functionals of a smooth density vs. estimating under the quadratic functionals, the plug in approach is strictly sub-optimal where the rates can be much improved (Bickel and Ritov, 1988; Donoho and Nussbaum, 1990).

In this paper, however, we prove that "estimating the Wasserstein-1 metric", is **not significantly easier** than "estimating under the Wasserstein-1 metric". Namely, the plug-in approach is minimax optimal up to a $\log\log(n)/\log(n)$ factor

$$\begin{aligned} \frac{\log\log(n \wedge m)}{\log(n \wedge m)} \cdot (n \wedge m)^{-\frac{\beta+1}{2\beta+d}} &\lesssim \inf_{\widetilde{T}_{m,n}} \sup_{\mu,\nu \in \mathcal{G}_\beta} \mathbf{E} \left| \widetilde{T}_{m,n} - W(\mu,\nu) \right| \\ &\leqslant \inf_{\widetilde{\mu}_m, \widetilde{\nu}_n} \sup_{\mu,\nu \in \mathcal{G}_\beta} \mathbf{E} \left| W(\widetilde{\mu}_m, \widetilde{\nu}_n) - W(\mu,\nu) \right| \lesssim (n \wedge m)^{-\frac{\beta+1}{2\beta+d}}, \end{aligned}$$

where $\mathcal{G}_\beta$ contains probability measures with densities in Hölder space with smoothness $\beta \in \mathbb{R}_{\geqslant 0}$. The result informs us that seeking other forms of estimators for $W(\mu,\nu)$ would only improve the rates logarithmically. The current result is in contrast with that in a forthcoming companion paper (Liang and Sadhanala, 2019), where we show that "estimating the adversarial losses" is **much easier** than "estimating under the adversarial losses", for a collection of integral probability metrics.

Remark that studying the Wasserstein metric and optimal transport for probability measures μ, ν with regularity condition has been an important topic in mathematics since Cafferalli's seminal result on regularity theory (Caffarelli, 1991, 1992). By studying the Monge-Ampére equation, Cafferalli showed that the Kantorovich potential satisfies specific regularity property, when μ, ν are Hölder smooth. In this paper, we follow the same Hölder smooth conditions on μ, ν, and study the statistical optimal rates for estimating $W(\mu,\nu)$, based on n-i.i.d samples.

1.1 Preliminaries

Let $\mathsf{C}^\beta(M) := \mathsf{C}^{\lfloor\beta\rfloor, \beta - \lfloor\beta\rfloor}(M)$ to be Hölder space with smoothness $\beta \in \mathbb{R}_{\geqslant 0}$.

$$\mathsf{C}^\beta(M) := \left\{ f : \Omega \to \mathbb{R} : \max_{|\alpha| \leqslant \lfloor\beta\rfloor} \sup_{x \in \Omega} |D^\alpha f| + \max_{|\alpha| = \lfloor\beta\rfloor} \sup_{x \neq y \in \Omega} \frac{|D^\alpha f(x) - D^\alpha f(y)|}{\|x - y\|^{\beta - \lfloor\beta\rfloor}} \leqslant M \right\} \tag{1.4}$$

where $\alpha = [\alpha_1, \ldots, \alpha_d] \in \mathbb{N}^d$ ranges over multi-indices, and $|\alpha| := \sum_{i=1}^d \alpha_i$. We only consider the bounded case with $\Omega = [0,1]^d$. The class of probability measures of interest is

$$\mathcal{G}_\beta := \left\{ \mu \ : \ \int_\Omega d\mu = 1, \mu \geqslant 0, \frac{d\mu}{dx} \in \mathsf{C}^\beta(M) \right\} \enspace . \tag{1.5}$$

The Wasserstein-1 metric is defined as

$$W_1(\mu,\nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{Y}} \|x - y\| d\pi \tag{1.6}$$

where $\Pi(\mu,\nu)$ denotes all coupling of probability measures μ, ν.

2 Optimal Rates for Estimating Wasserstein Metric

Theorem 1 (Minimax Rate)**.** *Consider $d \geqslant 2$ and the domain $\Omega = [0,1]^d$. Given m i.i.d. samples $X_1, \ldots, X_m$ from μ, and n i.i.d. samples from ν, then the minimax optimal rates for estimating $W(\mu, \nu)$ satisfies*

$$\frac{\log\log(n \wedge m)}{\log(n \wedge m)} \cdot (n \wedge m)^{-\frac{\beta+1}{2\beta+d}} \lesssim \inf_{\widetilde{T}_{m,n}} \sup_{\mu,\nu \in \mathcal{G}_\beta} \mathbf{E}\,|\widetilde{T}_{m,n} - W(\mu,\nu)| \lesssim (n \wedge m)^{-\frac{\beta+1}{2\beta+d}} \ , \tag{2.1}$$

where the μ, ν lies in $\mathcal{G}_\beta, \beta \geqslant 0$ as in (1.5) whose densities are β-Hölder smooth.

Remark 2.1. A few remarks are in order. First, we emphasize that the main technicality is in deriving the lower bound. We construct two composite/fuzzy hypotheses using delicate priors with matching $\log(n \wedge m)$ moments. However, the Wasserstein metric to estimate differs sufficiently under the null vs. under the alternative. Then we calculate the total variation metric directly on the posterior of data defined by the composite hypothesis, using a telescoping technique.

Second, as direct corollary, the following extension hold true. Suppose $\mu \in \mathcal{G}_{\beta_1}$ and $\nu \in \mathcal{G}_{\beta_2}$, then define $\beta := \beta_1 \wedge \beta_2$,

$$\frac{\log\log(n \wedge m)}{\log(n \wedge m)} \cdot (n \wedge m)^{-\frac{\beta+1}{2\beta+d}} \lesssim \inf_{\widetilde{T}_{m,n}} \sup_{\mu \in \mathcal{G}_{\beta_1}, \nu \in \mathcal{G}_{\beta_2}} \mathbf{E}\,|\widetilde{T}_{m,n} - W(\mu,\nu)| \lesssim (n \wedge m)^{-\frac{\beta+1}{2\beta+d}} \ . \tag{2.2}$$

A further direct implication is: when estimating the cost to transport a known measure $\mu \sim \mathrm{Unif}([0,1]^d)$ to an unknown ν based on $Y_1, \ldots, Y_n$, the result follows from setting $\beta_1 = \infty$ and $m = \infty$.

2.1 Proof of the Lower Bound

Without loss of generality, consider $m \geqslant n$. In the lower bound construction, we make use of the multi-resolution analysis. Denote $\mathsf{B}_q^{\beta,p}$ as the Besov space (Tribel, 1980; Donoho et al., 1996) with smoothness $\beta \in \mathbb{R}_{\geqslant 0}$, and $1 \leqslant p, q \leqslant \infty$,

$$\mathsf{B}_q^{\beta,p}(M) := \left\{ f(x) = \sum_{j=0}^{\infty} \sum_{k=0}^{2^{dj}-1} \theta_{jk} h_{jk}(x) : \left(\sum_{j=0}^{\infty} \left((2^{dj})^s \Big(\sum_{k=0}^{2^{dj}-1} |\theta_{jk}|^p \Big)^{1/p} \right)^q \right)^{1/q} \leqslant M, \text{ with } s = \frac{\beta}{d} + \frac{1}{2} - \frac{1}{p} \right\}$$

where $h_{jk}(x), x \in [0,1]^d$ is the wavelet basis. First, let us review some basic results on function spaces based on Tribel (1980); Donoho et al. (1996).

Proposition 2.1. *Under regularity conditions, the following equivalence holds between Besov space and Hölder space*

$$\mathsf{B}_\infty^{\beta,\infty} = \mathsf{C}^\beta, \textit{for } \beta \notin \mathbb{N} \tag{2.3}$$

In particular, when $\beta = 1$, $\mathsf{B}_\infty^{1,\infty} \supseteq \mathrm{Lip} \supseteq \mathsf{B}_1^{1,\infty}$.

Step 1: reduction to Besov space norm. Write $f_{jk} := \langle f, h_{jk} \rangle$, and $u_{jk} := \langle d\mu/dx, h_{jk} \rangle, v_{jk} := \langle d\nu/dx, v_{jk} \rangle$, we define the following integral probability metric as a surrogate

$$\begin{aligned}
d_{\mathsf{B}^{\gamma,p}_q}(\mu,\nu) &:= \sup_{f \in \mathsf{B}^{\gamma,p}_q} |\int f d\mu - \int f d\nu| \\
&= \sup_{f \in \mathsf{B}^{\gamma,p}_q} |\sum_{j \geqslant 0} \sum_{k=0}^{2^{dj}-1} f_{jk}(u_{jk} - v_{jk})| \\
&= \sup_{f \in \mathsf{B}^{\gamma,p}_q} |\sum_{j \geqslant 0} \|f_{j\cdot}\|_p \|u_{j\cdot} - v_{j\cdot}\|_{p_\star}| \\
&= \sup_{f \in \mathsf{B}^{\gamma,p}_q} |\sum_{j \geqslant 0} (2^{dj})^{\frac{\gamma}{d}+\frac{1}{2}-\frac{1}{p}} \|f_{j\cdot}\|_p \cdot (2^{-dj})^{\frac{\gamma}{d}+\frac{1}{2}-\frac{1}{p}} \|u_{j\cdot} - v_{j\cdot}\|_{p_\star}| \\
&= \left\{ \sum_{j \geqslant 0} \left[(2^{dj})^{\frac{\gamma}{d}+\frac{1}{2}-\frac{1}{p}} \|f_{j\cdot}\|_p \right]^q \right\}^{1/q} \left\{ \sum_{j \geqslant 0} \left[(2^{-dj})^{\frac{\gamma}{d}+\frac{1}{2}-\frac{1}{p}} \|u_{j\cdot} - v_{j\cdot}\|_{p_\star} \right]^{q_\star} \right\}^{1/q_\star} .
\end{aligned}$$

Take $p = q = \infty$ (in this case $p_\star = q_\star = 1$), we know

$$d_{\mathsf{B}^{\gamma,\infty}_\infty}(\mu,\nu) = \sum_{j \geqslant 0} (2^{-dj})^{\frac{\gamma}{d}+\frac{1}{2}} \sum_{k=0}^{2^j-1} |u_{jk} - v_{jk}|.$$

Take $p = \infty$, $q = 1$, we know

$$d_{\mathsf{B}^{\gamma,\infty}_\infty}(\mu,\nu) = \max_{j \geqslant 0} (2^{-dj})^{\frac{\gamma}{d}+\frac{1}{2}} \sum_{k=0}^{2^j-1} |u_{jk} - v_{jk}|.$$

Now the problem is related to estimation of weighted sum of ℓ_1 norm of the wavelet coefficients of the densities, in the following multiplicative sense

$$d_{\mathsf{B}^{\gamma,\infty}_1}(\mu,\nu) \leqslant W(\mu,\nu) \leqslant d_{\mathsf{B}^{\gamma,\infty}_\infty}(\mu,\nu) \ . \tag{2.4}$$

However, multiplicative equivalence is not enough for estimating $W(\mu,\nu)$. In our lower bound construction, we will show that for the hard instances of interest, equality holds.

Step 2: composite hypothesis testing. Next we are going to construct two priors on ν such that

$$|\mathop{\mathbf{E}}_{\nu \sim \mathcal{P}_0} W(\mu,\nu) - \mathop{\mathbf{E}}_{\nu \sim \mathcal{P}_1} W(\mu,\nu)| \tag{2.5}$$

are large, while one can not distinguish the following two distributions

$$p_0(Y_1, \dots Y_n) = \mathop{\mathbf{E}}_{\nu \sim \mathcal{P}_0} \left[\prod_{i=1}^n \frac{d\nu}{dx}(Y_i) \right], \ p_1(Y_1, \dots Y_n) = \mathop{\mathbf{E}}_{\nu \sim \mathcal{P}_1} \left[\prod_{i=1}^n \frac{d\nu}{dx}(Y_i) \right] \tag{2.6}$$

Here $\mathcal{P}_0, \mathcal{P}_1$ are two prior distributions on ν. Consider μ to be the same distribution under the null H_0 and the alternative H_1. Set

$$K \asymp \frac{\log n}{\log\log n}, \;\; \tau \asymp 1. \tag{2.7}$$

The choice will be clear in the later part of the proof. The prior construction is inspired from Lepski et al. (1999), where we borrow the following result.

Proposition 2.2. *For any given positive integer K and $\tau \in \mathbb{R}_{\geqslant 0}$, there exists two symmetric probability measures q_0 and q_1 on $[-\tau, \tau]$ such that*

$$\int_{-\tau}^{\tau} t^l q_0(dt) = \int_{-\tau}^{\tau} t^l q_1(dt), \quad l = 0, 1, \ldots, 2K; \tag{2.8}$$

$$\int_{-\tau}^{\tau} |t| q_1(dt) - \int_{-\tau}^{\tau} |t| q_0(dt) = 2\kappa \cdot K^{-1}\tau. \tag{2.9}$$

where κ is some constant depending on K only.

Now let's construct $\mathcal{P}_0$ and $\mathcal{P}_1$ as follows. Take $\mu \sim \mathrm{Unif}([0,1]^d)$. Choose $J \in \mathbb{N}_{\geqslant 0}$ such that $2^{dJ} \asymp n^{\frac{1}{1+2\beta/d}}$, first we are going to embed a parametrized class of densities into C^β

$$\frac{d\nu_\theta}{dx} := \mu(x) + \frac{1}{\sqrt{n}} \sum_{k=0}^{2^{dJ}-1} \theta_k h_{Jk}(x) \tag{2.10}$$

with $\theta_k \in [-\tau, \tau]$ for all k.

Let's verify $\nu_\theta \in \mathsf{B}_1^{\beta,\infty} \subseteq \mathsf{C}^\beta$ lies in the Hölder space. This follows since

$$\frac{1}{\sqrt{n}}|\theta_k| \leqslant (2^{dJ})^{-(\frac{\beta}{d}+\frac{1}{2})}, \quad \forall k. \tag{2.11}$$

And we have for any $\gamma \geqslant 0$

$$\begin{aligned} d_{\mathsf{B}_\infty^{\gamma,\infty}}(\mu, \nu_\theta) &:= (2^{-dJ})^{\frac{\gamma}{d}+\frac{1}{2}} \frac{1}{\sqrt{n}} \sum_{k=0}^{2^{dJ}-1} |\theta_k| \\ &= (2^{-dJ})^{\frac{\gamma}{d}+\frac{1}{2}} (2^{dJ})^{-(\frac{\beta}{d}+\frac{1}{2})} \sum_{k=0}^{2^{dJ}-1} |\theta_k| \\ &= (2^{-dJ})^{\frac{\beta+\gamma}{d}} \frac{1}{2^{dJ}} \sum_{k=0}^{2^{dJ}-1} |\theta_k|. \end{aligned}$$

It is easy to verify that

$$d_{\mathsf{B}_1^{\gamma,\infty}}(\mu, \nu_\theta) = (2^{-dJ})^{-\frac{\beta+\gamma}{d}} \cdot \frac{1}{2^{dJ}} \sum_{k \in [2^{dJ}]} |\theta_k| = d_{\mathsf{B}_\infty^{\gamma,\infty}}(\mu, \nu_\theta)$$

Therefore we must have for any $q \geqslant 1$, take $\gamma = 1$

$$W(\mu, \nu_\theta) = d_{\mathsf{B}_q^{1,\infty}}(\mu, \nu_\theta) = (2^{-dJ})^{-\frac{\beta+1}{d}} \cdot \frac{1}{2^{dJ}} \sum_{k \in [2^{dJ}]} |\theta_k|.$$

Step 3: polynomials and matching moments. Recall the collection of measures $\mathcal{S}_0 := \{\nu_\theta : \theta_k \sim q_0 \; i.i.d. \text{ for } k \in [2^{dJ}]\}$, and $\mathcal{P}_0$ can be viewed as an uniform prior over this set $\mathcal{S}_0$. Similar construction for $\mathcal{P}_1$ via q_1. Remark that due to the separation of support for wavelets, we have

$$\frac{d\nu_\theta}{dx} = \prod_{k=1}^{2^{dJ}} (1 + \theta_k n^{-1/2} h_{Jk}(x)) \ . \tag{2.12}$$

Therefore we know

$$\begin{aligned} p_0(Y_1, \ldots, Y_n) &= \mathop{\mathbf{E}}_{\theta \sim q_0^{\otimes 2^{dJ}}} \prod_{i=1}^{n} \frac{d\nu_\theta}{dx}(Y_i) = \mathop{\mathbf{E}}_{\theta \sim q_0^{\otimes 2^{dJ}}} \prod_{i=1}^{n} \prod_{k=1}^{2^{dJ}} (1 + \theta_k n^{-1/2} h_{Jk}(Y_i)) &(2.13)\\ &= \mathop{\mathbf{E}}_{\theta \sim q_0^{\otimes 2^{dJ}}} \prod_{k=1}^{2^{dJ}} \prod_{i=1}^{n} (1 + \theta_k n^{-1/2} h_{Jk}(Y_i)) &(2.14)\\ &= \prod_{k=1}^{2^{dJ}} \mathop{\mathbf{E}}_{\theta_k \sim q_0} \prod_{i=1}^{n} (1 + \theta_k n^{-1/2} h_{Jk}(Y_i)) \ . &(2.15) \end{aligned}$$

Let's analyze the polynomial in θ_k (and $h_{Jk}(Y_i)$) with degree at most n

$$\begin{aligned} f(\theta_k; h_{jk}(Y_1), \ldots, h_{jk}(Y_n)) &:= \prod_{i=1}^{n} (1 + \theta_k \frac{h_{Jk}(Y_i)}{\sqrt{n}}) &(2.16)\\ &= \sum_{l=0}^{n} \theta_k^l \frac{\sum_{i_1 < \ldots < i_l} h_{Jk}(Y_{i_1}) \ldots h_{Jk}(Y_{i_l})}{n^{l/2}} &(2.17)\\ &=: \sum_{l=0}^{n} \theta_k^l \frac{H_{JK}^{(l)}(Y_1, \ldots, Y_n)}{n^{l/2}} &(2.18) \end{aligned}$$

where $H_{JK}^{(l)}(Y_1, \ldots, Y_n)$ a sum of monomial of order l, i.e., $\binom{n}{l}$ terms with each of the form $h_{Jk}(Y_{i_1}) \ldots h_{Jk}(Y_{i_l})$. Denote $f^{[\leqslant K]}, f^{[>K]}$ to denote the corresponding truncated polynomial according to degree.

In this convenient notation, we know

$$p_0(Y_1, \ldots, Y_n) = \prod_{k \in [2^{dJ}]} \mathop{\mathbf{E}}_{\theta_k \sim q_0} f(\theta_k; h_{Jk}(Y_1), \ldots, h_{Jk}(Y_n)) \tag{2.19}$$

Later, we shall use the following properties of the polynomial f of degree at most n.

$$\forall \theta_k, \quad \int_{\mathcal{Y}^{\otimes n}} f(\theta_k; h_{Jk}(y_1), \ldots, h_{Jk}(y_n)) dy_1 \ldots dy_n = 1 \tag{2.20}$$

And the following property according to q_0 and q_1 constructed in Proposition 2.2: $\forall y_1, \ldots, y_n$

$$\begin{aligned} &\mathop{\mathbf{E}}_{\theta_k \sim q_1} f(\theta_k; h_{Jk}(y_1), \ldots, h_{Jk}(y_n)) - \mathop{\mathbf{E}}_{\theta_k \sim q_0} f(\theta_k; h_{Jk}(y_1), \ldots, h_{Jk}(y_n)) \\ &= \int_{[-\tau, \tau]} f^{[>2K]}(\theta_k; h_{Jk}(y_1), \ldots, h_{Jk}(y_n))(q_1 - q_0)(d\theta_k) \ . \end{aligned}$$

Step 4: total variation and telescoping.

$$
\begin{aligned}
\mathrm{TV}(p_1, p_0) &:= \frac{1}{2} \int_{\mathcal{Y}^{\otimes n}} \left| p_1(y_1, \ldots, y_n) - p_0(y_1, \ldots, y_n) \right| dy_1 \ldots dy_n \\
&= \frac{1}{2} \int_{\mathcal{Y}^{\otimes n}} \left| \prod_{k \in [2^{dJ}]} E_{\theta_k \sim q_1} f(\theta_k; h_{Jk}(y^{\otimes n})) - \prod_{k \in [2^{dJ}]} \underset{\theta_k \sim q_0}{\mathbf{E}} f(\theta_k; h_{Jk}(y^{\otimes n})) \right| dy_1 \ldots dy_n
\end{aligned}
$$

Claim the following telescoping lemma holds. The proof can be done through induction.

Proposition 2.3. *For all* $a_i, b_i \geqslant 0$,

$$
| \prod_{k \in [1,N]} a_k - \prod_{k \in [1,N]} b_k | \leqslant \sum_{i \in [1,N]} |a_i - b_i| \cdot \prod_{k \in [1,i)} b_k \cdot \prod_{k \in (i,N]} a_k \ . \tag{2.21}
$$

Define

$$
a_k(h_{Jk}(y_1), \ldots, h_{Jk}(y_n)) := E_{\theta_k \sim q_1} f(\theta_k; h_{Jk}(y^{\otimes n})) \tag{2.22}
$$

$$
b_k(h_{Jk}(y_1), \ldots, h_{Jk}(y_n)) := E_{\theta_k \sim q_0} f(\theta_k; h_{Jk}(y^{\otimes n})) \tag{2.23}
$$

Using the the above telescoping proposition, we have

$$
\mathrm{TV}(p_1, p_0) \leqslant \sum_{k \in [2^{dJ}]} \int |a_k - b_k| \cdot \prod_{k' \in [1,k)} b_{k'} \prod_{k'' \in (k,N]} a_{k''} dy^{\otimes n} \tag{2.24}
$$

$$
= \sum_{k \in [2^{dJ}]} \underset{\substack{\theta_{k'} \sim q_0, k' \in [1,k) \\ \theta_{k''} \sim q_1, k'' \in (k, 2^{dJ}]}}{\mathbf{E}} \ \underset{Y_1, \ldots, Y_n \sim \nu_{\theta_{-k}}}{\mathbf{E}} |a_k(h_{Jk}(Y_1), \ldots, h_{Jk}(Y_n)) - b_k(h_{Jk}(Y_1), \ldots, h_{Jk}(Y_n))| \tag{2.25}
$$

Let's analyze the term

$$
\underset{Y_1, \ldots, Y_n \sim \nu_{\theta_{-k}}}{\mathbf{E}} |a_k(h_{Jk}(Y_1), \ldots, h_{Jk}(Y_n)) - b_k(h_{Jk}(Y_1), \ldots, h_{Jk}(Y_n))|
$$

where $Y_1, \ldots Y_n$ i.i.d. sampled from a measure

$$
d\nu_{\theta_{-k}}/dx := 1 + \frac{1}{\sqrt{n}} \sum_{k' \neq k} \theta_{k'} h_{Jk'}(x) \ . \tag{2.26}
$$

Note that $\nu_{\theta_{-k}}$ agrees with the uniform measure μ on the domain associated with $h_{Jk}(x)$. Due to the separation of support for wavelet basis, we know the random variables

$$
h_{Jk}(Y_i) \tag{2.27}
$$

are only determined by $\nu_{\theta_{-k}}$ restricted to the domain of h_{Jk}. Hence for $Y_1, \ldots, Y_n \sim \nu_{\theta_{-k}}$,

$$
\begin{aligned}
&\underset{Y_1, \ldots, Y_n \sim \nu_{\theta_{-k}}}{\mathbf{E}} |a_k(h_{Jk}(Y_1), \ldots, h_{Jk}(Y_n)) - b_k(h_{Jk}(Y_1), \ldots, h_{Jk}(Y_n))| \\
&= \underset{Y_1, \ldots, Y_n \sim \mu}{\mathbf{E}} |a_k(h_{Jk}(Y_1), \ldots, h_{Jk}(Y_n)) - b_k(h_{Jk}(Y_1), \ldots, h_{Jk}(Y_n))| \ .
\end{aligned}
$$

Now one can directly bound the TV metric between the complex sum-product distribution p_0 and p_1 defined in (2.13),

$$2\mathrm{TV}(p_1, p_0) \leqslant \sum_{k=1}^{2^{dJ}} \mathop{\mathbf{E}}_{Y_1,\ldots,Y_n\sim\mu} |a_k(h_{Jk}(Y_1), \ldots, h_{Jk}(Y_n)) - b_k(h_{Jk}(Y_1), \ldots, h_{Jk}(Y_n))| \tag{2.28}$$

$$= \sum_{k=1}^{2^{dJ}} \int \left| \mathop{\mathbf{E}}_{\theta_k\sim q_1} f(\theta_k; h_{Jk}(y^{\otimes n})) - \mathop{\mathbf{E}}_{\theta_k\sim q_0} f(\theta_k; h_{Jk}(y^{\otimes n})) \right| dy_1 \ldots dy_n. \tag{2.29}$$

Step 5: ℓ_2 bound. In this section, we are going to bound, for a fixed k, the following expression using the properties of the q_1 and q_0 constructed with matching moments up to $2K$,

$$\int \left| \mathop{\mathbf{E}}_{\theta_k\sim q_1} f(\theta_k; h_{Jk}(y^{\otimes n})) - \mathop{\mathbf{E}}_{\theta_k\sim q_0} f(\theta_k; h_{Jk}(y^{\otimes n})) \right| dy_1 \ldots dy_n \quad .$$

First, observe the ℓ_2 bound

$$\int |g_1 - g_2| d\mu \leqslant \left(\int (g_1 - g_2)^2 d\mu \right)^{1/2} \tag{2.30}$$

Let's bound the ℓ_2 form

$$\int \left(\mathop{\mathbf{E}}_{\theta_k\sim q_1} f(\theta_k; h_{Jk}(y^{\otimes n})) - \mathop{\mathbf{E}}_{\theta_k\sim q_0} f(\theta_k; h_{Jk}(y^{\otimes n})) \right)^2 dy_1 \ldots dy_n \tag{2.31}$$

$$= \mathop{\mathbf{E}}_{\theta,\theta'\sim q_1} \int f(\theta; h_{Jk}(y^{\otimes n})) f(\theta'; h_{Jk}(y^{\otimes n})) dy^{\otimes n} + \mathop{\mathbf{E}}_{\omega,\omega'\sim q_0} \int f(\omega; h_{Jk}(y^{\otimes n})) f(\omega'; h_{Jk}(y^{\otimes n})) dy^{\otimes n}$$
$$- 2 \mathop{\mathbf{E}}_{\theta\sim q_1,\omega\sim q_0} \int f(\theta; h_{Jk}(y^{\otimes n})) f(\omega; h_{Jk}(y^{\otimes n})) dy^{\otimes n}$$

Note now each $f(\theta_k; h_{Jk}(y^{\otimes n}))f(\theta'; h_{Jk}(y^{\otimes n}))$ for fixed θ, θ' takes the following product form

$$f(\theta_k; h_{Jk}(y^{\otimes n})) f(\theta'; h_{Jk}(y^{\otimes n})) = \prod_{i=1}^{n} \left(1 + (\theta + \theta') \frac{h_{Jk}(Y_i)}{\sqrt{n}} + \theta\theta' \frac{h_{Jk}^2(Y_i)}{n} \right)$$

and

$$\int f(\theta; h_{Jk}(y^{\otimes n})) f(\theta'; h_{Jk}(y^{\otimes n})) dy^{\otimes n} = \left(1 + \theta\theta' \frac{\int h_{Jk}^2(y) dy}{n} \right)^n$$
$$= \left(1 + \theta\theta' \frac{1}{n} \right)^n \quad .$$

Therefore we have for (2.31)

$$
\begin{aligned}
(2.31) &= \mathop{\mathbf{E}}_{\theta,\theta'\sim q_1}\left[\left(1+\theta\theta'\frac{1}{n}\right)^n\right] + \mathop{\mathbf{E}}_{\omega,\omega'\sim q_0}\left[\left(1+\omega\omega'\frac{1}{n}\right)^n\right] - 2\mathop{\mathbf{E}}_{\theta\sim q_1,\omega\sim q_0}\left[\left(1+\theta\omega\frac{1}{n}\right)^n\right] \\
&= \sum_{l=1}^{\lfloor n/2\rfloor}\left(\mathop{\mathbf{E}}_{\theta,\theta'\sim q_1}[(\theta\theta')^{2l}] + \mathop{\mathbf{E}}_{\omega,\omega'\sim q_0}[(\omega\omega')^{2l}] - 2\mathop{\mathbf{E}}_{\theta\sim q_1,\omega\sim q_0}[(\theta\omega)^{2l}]\right)\frac{\binom{n}{2l}}{n^{2l}} \\
&= \sum_{l=1}^{\lfloor n/2\rfloor}\left(\left(\mathop{\mathbf{E}}_{q_1}[\theta^{2l}]\right)^2 + \left(\mathop{\mathbf{E}}_{q_0}[\theta^{2l}]\right)^2 - 2\mathop{\mathbf{E}}_{q_1}[\theta^{2l}]\mathop{\mathbf{E}}_{q_0}[\theta^{2l}]\right)\frac{\binom{n}{2l}}{n^{2l}}
\end{aligned}
$$

Recall the crucial property that for all $l \leqslant K$, we know

$$
\mathop{\mathbf{E}}_{\theta\sim q_1}[\theta^{2l}] = \mathop{\mathbf{E}}_{\theta\sim q_0}[\theta^{2l}] \quad\Rightarrow\quad \left(\mathop{\mathbf{E}}_{q_1}[\theta^{2l}]\right)^2 + \left(\mathop{\mathbf{E}}_{q_0}[\theta^{2l}]\right)^2 - 2\mathop{\mathbf{E}}_{q_1}[\theta^{2l}]\mathop{\mathbf{E}}_{q_0}[\theta^{2l}] = 0 \tag{2.32}
$$

therefore the above summation equals

$$
\begin{aligned}
(2.31) &= \sum_{l=K+1}^{\lfloor n/2\rfloor}\left(\left(\mathop{\mathbf{E}}_{q_1}[\theta^{2l}]\right)^2 + \left(\mathop{\mathbf{E}}_{q_0}[\theta^{2l}]\right)^2 - 2\mathop{\mathbf{E}}_{q_1}[\theta^{2l}]\mathop{\mathbf{E}}_{q_0}[\theta^{2l}]\right)\frac{\binom{n}{2l}}{n^{2l}} \\
&\leqslant \sum_{l=K+1}^{\lfloor n/2\rfloor} 4\tau^{4l}\frac{1}{(2l)!} \\
&\lesssim 4\frac{\tau^{4K}}{(2K)!}\exp(\tau^4) \quad .
\end{aligned}
$$

Assemble the two bounds, we have

$$
\begin{aligned}
&\int\left|\mathop{\mathbf{E}}_{\theta_k\sim q_1} f(\theta_k; h_{Jk}(y^{\otimes n})) - \mathop{\mathbf{E}}_{\theta_k\sim q_0} f(\theta_k; h_{Jk}(y^{\otimes n}))\right| dy_1\dots dy_n && (2.33) \\
&\leqslant 2\frac{\tau^{2K}}{\sqrt{(2K)!}}\exp(\tau^4/2) && (2.34)
\end{aligned}
$$

Step 6: combine all pieces. Now continuing (2.28), we have

$$
\begin{aligned}
2\mathrm{TV}(p_1,p_0) &\leqslant \sum_{k=1}^{2^{dJ}}\mathop{\mathbf{E}}_{Y_1,\dots,Y_n\sim\mu}|a_k(h_{Jk}(Y_1),\dots,h_{Jk}(Y_n)) - b_k(h_{Jk}(Y_1),\dots,h_{Jk}(Y_n))| && (2.35) \\
&= \sum_{k=1}^{2^{dJ}}\int\left|\mathop{\mathbf{E}}_{\theta_k\sim q_1} f(\theta_k; h_{Jk}(y^{\otimes n})) - \mathop{\mathbf{E}}_{\theta_k\sim q_0} f(\theta_k; h_{Jk}(y^{\otimes n}))\right| dy_1\dots dy_n && (2.36) \\
&\leqslant 2^{dJ}\cdot 2\frac{\tau^{2K}}{\sqrt{2K!}}\exp(\tau^4/2) \lesssim \exp(c\log n - K\log K) && (2.37)
\end{aligned}
$$

Therefore by taking $K = \frac{c}{2}\frac{\log n}{\log\log n}$, we know

$$
2\mathrm{TV}(p_1,p_0) \leqslant n^{-\frac{c}{2}\log n} \leqslant n^{-c/2}. \tag{2.38}
$$

We know by construction of the composite hypothesis

$$
\begin{aligned}
&|\mathop{\mathbf{E}}_{\nu_\theta\sim\mathcal{P}_0} d_{\mathsf{B}_q^{\gamma,\infty}}(\mu,\nu_\theta) - \mathop{\mathbf{E}}_{\nu_\theta\sim\mathcal{P}_1} d_{\mathsf{B}_q^{\gamma,\infty}}(\mu,\nu_\theta)| \\
&= (2^{-dJ})^{-\frac{\beta+\gamma}{d}} \cdot \left| \mathop{\mathbf{E}}_{\nu_\theta\sim\mathcal{P}_0} \left[\frac{1}{2^{dJ}} \sum_{k\in[2^{dJ}]} |\theta_k| \right] - \mathop{\mathbf{E}}_{\nu_\theta\sim\mathcal{P}_1} \left[\frac{1}{2^{dJ}} \sum_{k\in[2^{dJ}]} |\theta_k| \right] \right| \\
&= n^{-\frac{\beta+\gamma}{2\beta+d}} \cdot \left| \mathop{\mathbf{E}}_{\theta\sim q_0} [|\theta|] - \mathop{\mathbf{E}}_{\theta\sim q_1} [|\theta|] \right| \\
&\gtrsim n^{-\frac{\beta+\gamma}{2\beta+d}} \cdot 2\kappa K^{-1}\tau = n^{-\frac{\beta+\gamma}{2\beta+d}} \cdot \frac{\log\log(n)}{\log(n)} \quad .
\end{aligned}
$$

Therefore we have for any functional of θ, for any estimator based on n-i.i.d. samples

$$
\begin{aligned}
\sup_{\nu_\theta} \mathop{\mathbf{E}}_{\mathcal{D}_n\sim\theta} |\hat{T}_n - F(\theta)| &\geqslant \mathop{\mathbf{E}}_{\theta\sim Q_0} \mathbf{E}\, |\hat{T}_n - F(\theta)| \\
&\geqslant \mathop{\mathbf{E}}_{\theta\sim Q_0} \mathop{\mathbf{E}}_{\mathcal{D}_n\sim\theta} |\hat{T}_n - \mathop{\mathbf{E}}_{\theta\sim Q_0} F(\theta)| - \delta_{Q_0}
\end{aligned}
$$

where $\delta_{Q_0} := \mathbf{E}_{\theta\sim Q_0} |\, \mathbf{E}_{\theta\sim Q_0} F(\theta) - F_\theta|$. Here Q_0 is some prior distribution on θ. Repeat the same argument for Q_1, and by Le Cam's argument on two composite hypothesis

$$
\begin{aligned}
\sup_{\nu_\theta} \mathbf{E}\, |\hat{T}_n - F(\theta)| &\geqslant \frac{1}{2} \left(\mathop{\mathbf{E}}_{\theta\sim Q_0} \mathop{\mathbf{E}}_{\mathcal{D}_n\sim\theta} |\hat{T}_n - \mathop{\mathbf{E}}_{\theta\sim Q_0} F(\theta)| + \mathop{\mathbf{E}}_{\theta\sim Q_1} \mathop{\mathbf{E}}_{\mathcal{D}_n\sim\theta} |\hat{T}_n - \mathop{\mathbf{E}}_{\theta\sim Q_1} F(\theta)| \right) - \frac{\delta_{Q_0}+\delta_{Q_1}}{2} \\
&= \frac{1}{2} \left(\mathop{\mathbf{E}}_{\mathcal{D}_n\sim p_0} |\hat{T}_n - \mathop{\mathbf{E}}_{\theta\sim Q_0} F(\theta)| + \mathop{\mathbf{E}}_{\mathcal{D}_n\sim p_1} |\hat{T}_n - \mathop{\mathbf{E}}_{\theta\sim Q_1} F(\theta)| \right) - \frac{\delta_{Q_0}+\delta_{Q_1}}{2} \\
&\geqslant \frac{|\, \mathbf{E}_{\theta\sim Q_0} F(\theta) - \mathbf{E}_{\theta\sim Q_1} F(\theta)|}{4} \big(P_0(T=1) + P_1(T=0) \big) - \frac{\delta_{Q_0}+\delta_{Q_1}}{2} \\
&\geqslant \frac{|\, \mathbf{E}_{\theta\sim Q_0} F(\theta) - \mathbf{E}_{\theta\sim Q_1} F(\theta)|}{4} \int p_0(y^{\otimes n}) \wedge p_1(y^{\otimes n}) dy^{\otimes n} - \frac{\delta_{Q_0}+\delta_{Q_1}}{2} \\
&= \frac{|\, \mathbf{E}_{\theta\sim Q_0} F(\theta) - \mathbf{E}_{\theta\sim Q_1} F(\theta)|}{4} (1 - d_{TV}(p_0,p_1)) - \frac{\delta_{Q_0}+\delta_{Q_1}}{2}
\end{aligned}
$$

where $p_i(y^{\otimes n}) = \int Pr(y^{\otimes n}|\theta) Q_i(d\theta)$, for $i = 0, 1$. Here the test $T = 1$ if and only if $\hat{T}_n$ is closer to $\mathbf{E}_{\theta\sim Q_1} F(\theta)$. In our case, for any $q \geqslant 1$

$$
F(\theta) := W(\mu,\nu) = d_{\mathsf{B}_q^{1,\infty}}(\mu,\nu_\theta) = (2^{-dJ})^{-\frac{\beta+1}{d}} \left[\frac{1}{2^{dJ}} \sum_{k\in[2^{dJ}]} |\theta_k| \right]
$$

then

$$
\begin{aligned}
|\mathop{\mathbf{E}}_{\theta\sim Q_0} F(\theta) - \mathop{\mathbf{E}}_{\theta\sim Q_1} F(\theta)| &= |\mathop{\mathbf{E}}_{\nu_\theta\sim\mathcal{P}_0} d_{\mathsf{B}_q^{1,\infty}}(\mu,\nu_\theta) - \mathop{\mathbf{E}}_{\nu_\theta\sim\mathcal{P}_1} d_{\mathsf{B}_q^{1,\infty}}(\mu,\nu_\theta)| \\
&\gtrsim n^{-\frac{\beta+1}{2\beta+d}} \cdot \frac{\log\log(n)}{\log(n)} \\
1 - d_{TV}(p_0,p_1) &\geqslant 1 - n^{-c/2} \\
\frac{\delta_{Q_0}+\delta_{Q_1}}{2} &\lesssim n^{-\frac{\beta+1}{2\beta+d}} \frac{1}{\sqrt{2^{dJ}}} \ll n^{-\frac{\beta+1}{2\beta+d}} \cdot \frac{\log\log(n)}{\log(n)} \quad .
\end{aligned}
$$

Therefore we have

$$\inf_{\widehat{T}_n} \sup_{\nu \in \mathsf{C}^\beta} \mathbf{E}\, |\widehat{T}_n - W(\mu,\nu)| \gtrsim n^{-\frac{\beta+1}{2\beta+d}} \cdot \frac{\log\log(n)}{\log(n)} \ . \tag{2.39}$$

2.2 Proof of the Upper Bound

The upper bound can be obtained through similar derivations as in Liang (2018); Singh et al. (2018); Weed and Berthet (2019). We include here for completeness.

The estimator is of the plug-in form, with

$$W(\widetilde{\mu}_m, \widetilde{\nu}_n) := \sup_{f \in \mathrm{Lip}(1)} |\int f d\widetilde{\mu}_m - \int f d\widetilde{\nu}_n| \tag{2.40}$$

where $\widetilde{\mu}_m$, and $\widetilde{\nu}_n$ are smoothed empirical measures based on truncation on Wavelets. It is clear that

$$|W(\widetilde{\mu}_m, \widetilde{\nu}_n) - W(\mu,\nu)| \leqslant \sup_{f \in \mathrm{Lip}(1)} |\int f d\widetilde{\mu}_m - \int f d\mu| + \sup_{f \in \mathrm{Lip}(1)} |\int f d\widetilde{\nu}_n - \int f d\nu|. \tag{2.41}$$

Now let's bound $\sup_{f \in \mathrm{Lip}(1)} |\int f d\widetilde{\nu}_n - \int f d\nu|$ via expanding under the Wavelet basis. Denote $\widehat{\mathbf{E}}[h_{jk}] := 1/n \sum_{i=1}^n h_{jk}(Y_i)$, the smoothed empirical estimate $\widetilde{\nu}_n$ is defined

$$\frac{d\widetilde{\nu}_n}{dx} := \sum_{j=0}^{J} \sum_{k=0}^{2^{dj}-1} \widehat{\mathbf{E}}[h_{jk}] h_{jk}(x) \ . \tag{2.42}$$

Expand $f(x) = \sum_{j \geqslant 0} \sum_{k=0}^{2^{dj}-1} f_{jk} h_{jk}(x)$, we have

$$\begin{aligned}
&\sup_{f \in \mathrm{Lip}(1)} |\int f d\widetilde{\nu}_n - \int f d\nu| \leqslant \sup_{f \in \mathsf{B}^{1,\infty}_{\infty}} |\int f d\widetilde{\nu}_n - \int f d\nu| \\
&= \sup_{f \in \mathsf{B}^{1,\infty}_{\infty}} |\sum_{j \geqslant 0}^{J} \sum_{k=0}^{2^{dj}-1} f_{jk}(\widehat{\mathbf{E}}[h_{jk}] - \mathbf{E}[h_{jk}])| + \sup_{f \in \mathsf{B}^{1,\infty}_{\infty}} |\sum_{j > J} \sum_{k=0}^{2^{dj}-1} f_{jk}\, \mathbf{E}[h_{jk}]|
\end{aligned}$$

For the first term, since $f \in \mathsf{B}^{1,\infty}_{\infty} \Rightarrow \forall j,k,\ |f_{jk}| \leqslant (2^{-dj})^{\frac{1}{d}+\frac{1}{2}}$

$$\begin{aligned}
&\mathbf{E} \sup_{f \in \mathsf{B}^{1,\infty}_{\infty}} |\sum_{j \geqslant 0}^{J} \sum_{k=0}^{2^{dj}-1} f_{jk}(\widehat{\mathbf{E}}[h_{jk}] - \mathbf{E}[h_{jk}])| \leqslant \sum_{j \geqslant 0}^{J} (2^{-dj})^{\frac{1}{d}+\frac{1}{2}} \sum_{k=0}^{2^{dj}-1} \mathbf{E}\, |\widehat{\mathbf{E}}[h_{jk}] - \mathbf{E}[h_{jk}]| \\
&\leqslant \sum_{j \geqslant 0}^{J} (2^{-dj})^{\frac{1}{d}+\frac{1}{2}} \sum_{k=0}^{2^{dj}-1} (\mathbf{E}\, |\widehat{\mathbf{E}}[h_{jk}] - \mathbf{E}[h_{jk}]|^2)^{1/2} \\
&\precsim \sum_{j \geqslant 0}^{J} (2^{-dj})^{\frac{1}{d}+\frac{1}{2}} 2^{dj} \frac{1}{\sqrt{n}} \asymp \frac{1}{\sqrt{n}} (2^{dJ})^{\frac{1}{2}-\frac{1}{d}}
\end{aligned}$$

for $d \geqslant 2$.

For the second term, recall $\mathbf{E}_{Y\sim\nu}[h_{jk}(Y)] = \langle d\nu/dx, h_jk\rangle =: v_{jk}$. Due to the fact that

$$d\nu/dx \in \mathsf{C}^{\beta} \in \mathsf{B}^{\beta,\infty}_{\infty} \Rightarrow \forall j,k,\ |v_{jk}| \leqslant (2^{-dj})^{\frac{\beta}{d}+\frac{1}{2}} \tag{2.43}$$

$$f \in \mathsf{B}^{1,\infty}_{\infty} \Rightarrow \forall j,k,\ |f_{jk}| \leqslant (2^{-dj})^{\frac{1}{d}+\frac{1}{2}} \tag{2.44}$$

$$\begin{aligned}
&\mathbf{E} \sup_{f\in\mathsf{B}^{1,\infty}_{\infty}} |\sum_{j>J} \sum_{k=0}^{2^{dj}-1} f_{jk}\, \mathbf{E}[h_{jk}]| = \mathbf{E} \sup_{f\in\mathsf{B}^{1,\infty}_{\infty}} |\sum_{j>J} \sum_{k=0}^{2^{dj}-1} f_{jk} v_{jk}| \\
&\leqslant \sum_{j>J} \sum_{k=0}^{2^{dj}-1} (2^{-dj})^{\frac{1}{d}+\frac{1}{2}} (2^{-dj})^{\frac{\beta}{d}+\frac{1}{2}} \\
&\leqslant (2^{dJ})^{-\frac{\beta+1}{d}}
\end{aligned}$$

Balancing the two terms, we have

$$\sup_{\nu\in\mathcal{G}_\beta} \sup_{f\in \mathrm{Lip}(1)} |\int f d\widetilde{\nu}_n - \int f d\nu| \lesssim \frac{1}{\sqrt{n}} (2^{dJ})^{\frac{1}{2}-\frac{1}{d}} + (2^{dJ})^{-\frac{\beta+1}{d}} \tag{2.45}$$

$$\asymp n^{-\frac{\beta+1}{2\beta+d}}, \quad \text{with } 2^{dJ} \asymp n^{\frac{1}{2\beta/d+1}} \ . \tag{2.46}$$

Put everything together, we know

$$\mathbf{E}\, |W(\widetilde{\mu}_m, \widetilde{\nu}_n) - W(\mu,\nu)| \leqslant (n \wedge m)^{-\frac{\beta+1}{2\beta+d}}. \tag{2.47}$$

References

Peter J Bickel and Yaacov Ritov. Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 381–393, 1988.

Luis A Caffarelli. Some regularity properties of solutions of monge ampere equation. *Communications on pure and applied mathematics*, 44(8-9):965–969, 1991.

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

Luis A Caffarelli. Boundary regularity of maps with convex potentials–ii. *Annals of mathematics*, 144(3):453–496, 1996.

David L Donoho and Michael Nussbaum. Minimax quadratic estimation of a quadratic functional. *Journal of Complexity*, 6(3):290–323, 1990.

David L Donoho, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard. Density estimation by wavelet thresholding. *The Annals of Statistics*, pages 508–539, 1996.

Richard Mansfield Dudley. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

Oleg Lepski, Arkady Nemirovski, and Vladimir Spokoiny. On estimation of the l r norm of a regression function. *Probability theory and related fields*, 113(2):221–253, 1999.

Tengyuan Liang. How well can generative adversarial networks learn densities: A nonparametric view. *arXiv preprint arXiv:1712.08244*, 2017.

Tengyuan Liang. On how well generative adversarial networks learn densities: Nonparametric and parametric results. *arXiv preprint arXiv:1811.03179*, 2018.

Tengyuan Liang and Veeranjaneyulu Sadhanala. On minimax optimality of estimating the adversarial losses. Technical report, University of Chicago, 2019.

Michael Nussbaum et al. Asymptotic equivalence of density estimation and gaussian white noise. *The Annals of Statistics*, 24(6):2399–2430, 1996.

Shashank Singh, Ananya Uppal, Boyue Li, Chun-Liang Li, Manzil Zaheer, and Barnabás Póczos. Nonparametric density estimation under adversarial losses. In *Advances in Neural Information Processing Systems*, pages 10225–10236, 2018.

H Tribel. Theory of interpolation, functional spaces and differential operators. 1980.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *arXiv preprint arXiv:1707.00087*, 2017.

Jonathan Weed and Quentin Berthet. Estimation of smooth densities in wasserstein distance. *arXiv preprint arXiv:1902.01778*, 2019.