Estimating Certain Integral Probability Metric (IPM) Is as Hard as Estimating under the IPM

Tengyuan Liang*[1]

[1]University of Chicago

Abstract

We study the minimax optimal rates for estimating a range of Integral Probability Metrics (IPMs) between two unknown probability measures, based on n independent samples from them. Curiously, we show that estimating the IPM itself between probability measures, is not significantly easier than estimating the probability measures under the IPM. 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 Integral Probability Metrics (IPMs) between probability measures based on samples. IPMs are widely used in both statistics and machine learning, with applications in nonparametric two-sample tests [22; 10], inferring the transportation cost (the Wasserstein-1 metric) from one set of samples to another [21; 19], and with more recent appearances in rigorous investigations on the generative adversarial networks (GANs) [1; 17; 14; 20; 24; 3].

Let μ, ν be two probability measures supported on $\Omega = [0,1]^d$, and $d_{\mathcal{F}}(\mu,\nu)$ denote a certain IPM between them induced by a set of functions $\mathcal{F}$, defined as

$$d_{\mathcal{F}}(\mu,\nu) := \sup_{f\in\mathcal{F}} \left| \int_\Omega f d\mu - \int_\Omega f d\nu \right| \ . \tag{1.1}$$

Consider that $X_1, \dots X_m$ 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 $d_{\mathcal{F}}(\mu,\nu)$ based on $\{X_i\}_{i=1}^m, \{Y_j\}_{j=1}^n$, for some class of probability measures $\mathcal{G}$ of interest

$$\inf_{\widetilde{T}_{m,n}} \sup_{\mu,\nu\in\mathcal{G}} \mathbf{E} \left| \widetilde{T}_{m,n} - d_{\mathcal{F}}(\mu,\nu) \right| \ . \tag{1.2}$$

It turns out that using the empirical measure $\widehat{\nu}_n$ to estimate is a bad idea when $\mathcal{F}$ is complex enough, regardless of how simple $\mathcal{G}$ is. To see this, let's consider a simple case with $\mathcal{F} = \mathrm{Lip}(1)$. In

*Liang gratefully acknowledges support from the George C. Tiao Fellowship. The paper was previously titled "On the minimax optimality of estimating Wasserstein metric" [15]. The previous version is no longer intended for publication.

such a case, $d_{\mathcal{F}}$ reduces to the Wasserstein-1 metric W (1.10). Due to a result by Dudley [8], even for infinitely smooth $\mathcal{G} = \{\mathrm{Unif}(\Omega)\}$ and $d \geqslant 2$,

$$\sup_{\nu \in \mathcal{G}} |W(\mu, \widehat{\nu}_n) - W(\mu, \nu)| \asymp n^{-\frac{1}{d}} \ . \tag{1.3}$$

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

A related yet different problem studied in the current literature is estimating a probability measure under certain IPMs [22; 25; 14; 20; 26], in the following sense

$$\inf_{\widetilde{\nu}_n} \sup_{\nu \in \mathcal{G}} \mathbf{E}\, d_{\mathcal{F}}(\widetilde{\nu}_n, \nu) \ . \tag{1.4}$$

The two problems are closely related: "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 former, since

$$\mathbf{E}\, |d_{\mathcal{F}}(\widetilde{\mu}_m, \widetilde{\nu}_n) - d_{\mathcal{F}}(\mu, \nu)| \leqslant \mathbf{E}\, d_{\mathcal{F}}(\widetilde{\mu}_m, \mu) + \mathbf{E}\, d_{\mathcal{F}}(\widetilde{\nu}_n, \nu) \ .$$

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 [2; 6; 9]. In recent practical applications such as GANs, one is curious to understand if evaluating and inferring how well we do in terms of learning the probability measure, could be simpler than learning the measure itself [18; 16].

In this paper, however, we prove that "estimating the IPMs," is **not significantly easier** than "estimating under the IPMs," for a wide range of measures and metrics. Specifically, the plug-in approach is minimax optimal up to a $\log\log(n)/\log(n)$ factor

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

Here $\mathcal{G}_\beta$ contains probability measures with densities in the Hölder space with smoothness $\beta \in \mathbb{R}_{\geqslant 0}$, and the IPMs are induced by $\mathcal{F}_\gamma$, the Hölder space with smoothness $\gamma \in \mathbb{R}_{\geqslant 0}$, with $\gamma < d/2$. Note that when $\gamma \geqslant d/2$, the parametric rate $n^{-1/2}$ is attainable. The result informs us that (1) seeking for other forms of estimators for $d_{\mathcal{F}_\gamma}(\mu, \nu)$ would only improve the rates logarithmically, and (2) estimating the IPM between two measures is fundamentally just as hard as estimating the measure under the IPM.

1.1 Preliminaries

We introduce the notations used in the paper. For a function $f : \mathbb{R}^d \to \mathbb{R}$ and $p \geqslant 1$, $\|f\|_{L_p}$ denotes the L_p norm w.r.t. the Lebesgue measure. For a finite dimensional vector θ and $q \geqslant 1$, $\|\theta\|_q$ is the vector L_q norm, and $\|\theta\|$ is the L_2 norm. For an integer K, $[K] := \{0, 1, \ldots, K-1\}$.

Let $\mathsf{C}^\beta := \mathsf{C}^{\lfloor\beta\rfloor,\beta-\lfloor\beta\rfloor}$ to be Hölder space with smoothness $\beta > 0$.

$$\mathsf{C}^\beta := \left\{ f : \Omega \to \mathbb{R} : \max_{|\alpha|\leqslant\lfloor\beta\rfloor} \sup_{x\in\Omega} |D^\alpha f| + \max_{|\alpha|=\lfloor\beta\rfloor} \sup_{x\neq y\in\Omega} \frac{|D^\alpha f(x) - D^\alpha f(y)|}{\|x-y\|^{\beta-\lfloor\beta\rfloor}} < \infty \right\} \tag{1.5}$$

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 $\Omega = [0,1]^d$.

Define the forward difference operator, given a function $f : \mathbb{R}^d \to \mathbb{R}$, for every $h \in \mathbb{R}^d$

$$\begin{aligned} \Delta_h f(x) &:= f(x+h) - f(x), \\ \Delta_h^m f(x) &:= \Delta_h(\Delta_h^{m-1} f(x)),\ m \geqslant 2. \end{aligned}$$

Let $1 \leqslant p, q \leqslant \infty$ and $\beta > 0$, the Besov-Lipschitz space semi-norm $\|\cdot\|'_{\mathsf{B}_q^{\beta,p}}$ is defined in the following way [12, Chapter 17, Proposition 17.21].

$$\|f\|'_{\mathsf{B}_q^{\beta,p}} := \left(\sum_{j=0}^\infty \left((2^j)^\beta \sup_{\|h\|\leqslant 1/2^j} \|\Delta_h^{\lfloor\beta\rfloor+1} f\|_{L_p} \right)^q \right)^{1/q}. \tag{1.6}$$

Wavelets are used to provide an equivalent characterization of the Besov spaces, if the basis $\{h_{jk}, j \in \mathbb{N}, 0 \leqslant k < 2^{jd}\}$ satisfies certain regularity conditions [11, Chapter 9, Theorem 9.2]. For function $f(x) = \sum_{j=0}^\infty \sum_{k<2^{dj}} \theta_{jk} h_{jk}(x)$, define the Besov space norm in terms of the wavelet coefficients [26]

$$\|f\|_{\mathsf{B}_q^{\beta,p}} := \left(\sum_{j=0}^\infty \left((2^j)^\beta (2^{jd})^{\frac{1}{2}-\frac{1}{p}} \|\theta_{j\cdot}\|_p \right)^q \right)^{1/q}. \tag{1.7}$$

In this paper, we assume such regularity conditions throughout so that the Besov space norms $\|\cdot\|'_{\mathsf{B}_q^{\beta,p}}$ and $\|\cdot\|_{\mathsf{B}_q^{\beta,p}}$ are equivalent. We refer the readers to [26, Appendix C] and [5, Chapter 2.12] for details on the regularity conditions that we assume on the wavelets.

Besov spaces subsume Hölder spaces as special cases ($p = q = \infty$), in the following sense [23; 7; 11]: under regularity conditions, the following equivalence holds between the Besov space and Hölder space

$$\mathsf{B}_\infty^{\beta,\infty} = \mathsf{C}^\beta, \text{for } \beta \notin \mathbb{N}.$$

In particular, when $\beta = 1$, $\mathsf{B}_\infty^{\beta,\infty}$ is called the Zygmund space, which contains the Lipschitz space $\mathsf{B}_\infty^{1,\infty} \supseteq \text{Lip} \supseteq \mathsf{B}_1^{1,\infty}$.

Now we are ready to formally state the parameter spaces, and the IPMs to study.

Parameter Spaces. For some $M > 0$, the class of probability measures of interest is

$$\mathcal{G}_\beta := \left\{ \mu \ : \ \int_\Omega d\mu = 1, \mu \geqslant 0, \frac{d\mu}{dx} \in \mathsf{B}_\infty^{\beta,\infty}(M) \right\}, \tag{1.8}$$

with

$$\mathsf{B}_\infty^{\beta,\infty}(M) := \left\{ f \ : \ \|f\|_{\mathsf{B}_\infty^{\beta,\infty}} \leqslant M \right\}.$$

Again, for non-integer β, we are considering densities that are Hölder smooth.

Integral Probability Metric. The class of IPMs considered is induced by the Besov space, for some $\gamma > 0$

$$\mathcal{F}_\gamma := \mathsf{B}_\infty^{\gamma,\infty}(1) \ ,$$
$$d_{\mathcal{F}_\gamma}(\mu,\nu) = \sup_{f\in\mathcal{F}_\gamma} \left| \int f d\mu - \int f d\nu \right| \ . \tag{1.9}$$

As a special case for the IPMs, the Wasserstein-1 metric (for measures supported on bounded Ω) is

$$W(\mu,\nu) := \sup_{f\in \mathrm{Lip}(1)} \left| \int f d\mu - \int f d\nu \right| \ . \tag{1.10}$$

2 Optimal Rates for Estimating IPMs

Theorem 1 (Minimax Rate)**.** *Consider 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 $Y_1,\ldots,Y_n$ from ν. Then the minimax optimal rate for estimating $d_{\mathcal{F}_\gamma}(\mu,\nu)$ satisfies*

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

Here μ,ν lie in $\mathcal{G}_\beta, \beta\in\mathbb{R}_{\geqslant 0}$ as in (1.8) whose densities are β-Hölder smooth. The function class $\mathcal{F}_\gamma, \gamma\in\mathbb{R}_{\geqslant 0}$ for the metric is defined in (1.9), with $\gamma < d/2$.

Remark 2.1. Here the β quantifies the regularity of the measures, and γ quantifies the regularity of the metrics.

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 IPMs to estimate differs sufficiently under the null vs. the alternative. Then we calculate the Total Variation (TV) metric directly on the posterior of data samples defined by the composite hypothesis, using some telescoping techniques involving sum-products. The transparent technique could be of independent interest in handling TV-type calculations in proving lower bounds.

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

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

Third, the $\gamma < d/2$ condition is effectively equivalent to that $\mathcal{F}_\gamma$ is beyond the Donsker's class. This is the complex regime since within the Donsker's class $\gamma \geqslant d/2$, the parametric rate $n^{-1/2}$ is attainable [10; 14; 20].

2.1 Proof of the Lower Bound

Without the loss of generality, consider the case when $m \geqslant n$. The lower bound construction is divided into six logical steps, for better organization. We make use of multi-resolution analysis in the construction.

Step 1: reduction to Besov space semi-norm. For any $p \geqslant 1$, define $p_\star \geqslant 1$ such that $1/p_\star + 1/p = 1$. For simplicity, define Radon-Nikodym derivative of measure μ w.r.t. the Lebesgue measure as $\rho_\mu(x) := d\mu/dx$. Define explicitly the wavelet coefficients $f_{jk} := \langle f, h_{jk}\rangle$, and $u_{jk} := \langle d\mu/dx, h_{jk}\rangle, v_{jk} := \langle d\nu/dx, h_{jk}\rangle$. Under such notations, the integral probability metric reduces to the following

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

Take $p = q = \infty$ (in this case $p_\star = q_\star = 1$), we know that the IPM can be regarded as a type of Besov space semi-norm

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

Step 2: composite hypothesis and prior construction. Next we are going to construct two priors on ν such that the difference

$$|\mathop{\mathbf{E}}_{\nu\sim\mathcal{P}_0} d_{\mathcal{F}_\gamma}(\mu,\nu) - \mathop{\mathbf{E}}_{\nu\sim\mathcal{P}_1} d_{\mathcal{F}_\gamma}(\mu,\nu)|$$

is large, while at the same time one can not distinguish the following two distributions

$$p_0(Y_1,\dots Y_n) = \mathop{\mathbf{E}}_{\nu\sim\mathcal{P}_0}\left[\prod_{i=1}^n \rho_\nu(Y_i)\right],\ p_1(Y_1,\dots Y_n) = \mathop{\mathbf{E}}_{\nu\sim\mathcal{P}_1}\left[\prod_{i=1}^n \rho_\nu(Y_i)\right]. \tag{2.2}$$

Here $\mathcal{P}_0, \mathcal{P}_1$ are two prior distributions on ν which we will construct. Consider μ to be the same distribution under the null H_0 and the alternative H_1. Set two values K, τ to be used in the construction

$$K \asymp \frac{\log n}{\log\log n},\ \tau \asymp 1. \tag{2.3}$$

The choice will be apparent in the latter part of the proof. The following prior construction is inspired by [13], where they study the estimation of functionals under the Gaussian white noise model. This prior was also used in [4] for studying non-smooth functional estimation in Gaussian sequence models.

Proposition 2.1 ([13], Proposition 4.2)**.** *For any given positive integer K and $\tau \in \mathbb{R}_{\geqslant 0}$, there exist two symmetric probability measures q_0 and q_1 on $[-\tau, \tau]$ such that*

$$\int_{-\tau}^{\tau} t^l q_0(dt) = \int_{-\tau}^{\tau} t^l q_1(dt), \quad l = 0, 1, \ldots, 2K, \tag{2.4}$$

$$\int_{-\tau}^{\tau} |t| q_1(dt) - \int_{-\tau}^{\tau} |t| q_0(dt) = 2\kappa \cdot K^{-1}\tau, \tag{2.5}$$

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 $J \in \mathbb{N}$ such that $2^{dJ} \asymp n^{\frac{1}{1+2\beta/d}}$, first we embed a parametrized class of densities into $\mathcal{G}_\beta$

$$\frac{d\nu_\theta}{dx} := \mu(x) + \frac{1}{\sqrt{n}} \sum_{k=0}^{2^{dJ}-1} \theta_k h_{Jk}(x) \tag{2.6}$$

with each $\theta_k \in [-\tau, \tau]$ for all k. Now we show that the construction lies inside the space of interest, i.e., $\nu_\theta \in \mathcal{G}_\beta$. First observe that for the wavelet basis that satisfy the regularity condition $\int_\Omega h_{jk} d\mu = 0$, we have $\int_\Omega d\nu_\theta = 1$ and $\|d\nu_\theta/dx\|_{L_\infty} \geqslant 1 - \sqrt{2^{dJ}/n} > 0$. Hence ν_θ is a valid probability measure. Let's then verify that the density $\rho_{\nu_\theta} \in \mathsf{B}_\infty^{\beta,\infty}$ lies in the Besov space, since

$$\frac{1}{\sqrt{n}} |\theta_k| \leqslant (2^{dJ})^{-(\frac{\beta}{d} + \frac{1}{2})}, \quad \forall k. \tag{2.7}$$

For any $\gamma \geqslant 0$, it is then easy to verify via Step 1 that

$$\begin{aligned} d_{\mathcal{F}_\gamma}(\mu, \nu_\theta) &:= (2^{-dJ})^{\frac{\gamma}{d} + \frac{1}{2}} \frac{1}{\sqrt{n}} \sum_{k=0}^{2^{dJ}-1} |\theta_k| \\ &= (2^{-dJ})^{\frac{\gamma}{d} + \frac{1}{2}} (2^{dJ})^{-(\frac{\beta}{d} + \frac{1}{2})} \sum_{k=0}^{2^{dJ}-1} |\theta_k| \\ &= (2^{-dJ})^{\frac{\beta+\gamma}{d}} \frac{1}{2^{dJ}} \sum_{k=0}^{2^{dJ}-1} |\theta_k|. \end{aligned} \tag{2.8}$$

Making use of the probability measures q_0 and q_1 on $[-\tau, \tau]$ claimed by Proposition 2.1, we define a collection of measures

$$\mathcal{S}_0 := \{\nu_\theta : \theta_k \sim q_0 \ i.i.d. \text{ for each } k \in [2^{dJ}]\}.$$

Then $\mathcal{P}_0$ can be viewed as an uniform prior over this set $\mathcal{S}_0$. Similar construction holds for $\mathcal{P}_1$ via q_1.

Step 3: polynomials and matching moments. Remark that due to the separation of support for wavelets (localized property), i.e., $h_{Jk}(x) h_{Jk'}(x) = 0$ for $k \neq k'$, we have the equivalent expression as in (2.6)

$$\frac{d\nu_\theta}{dx} = \prod_{k=0}^{2^{dJ}-1} \left(1 + \theta_k n^{-1/2} h_{Jk}(x)\right) . \tag{2.9}$$

Use $\theta \sim q_0^{\otimes 2^{dJ}}$ to denote that $\theta_k \sim q_0$ i.i.d. for all $k \in [2^{dJ}]$, we know

$$\begin{aligned} p_0(Y_1,\dots,Y_n) &= \mathop{\mathbf{E}}_{\theta\sim q_0^{\otimes 2^{dJ}}} \prod_{i=1}^n \rho_\theta(Y_i) = \mathop{\mathbf{E}}_{\theta\sim q_0^{\otimes 2^{dJ}}} \prod_{i=1}^n \prod_{k=0}^{2^{dJ}-1} (1+\theta_k n^{-1/2} h_{Jk}(Y_i)) \\ &= \mathop{\mathbf{E}}_{\theta\sim q_0^{\otimes 2^{dJ}}} \prod_{k=0}^{2^{dJ}-1} \prod_{i=1}^n (1+\theta_k n^{-1/2} h_{Jk}(Y_i)) \quad \text{by (2.9)} \\ &= \prod_{k=1}^{2^{dJ}} \mathop{\mathbf{E}}_{\theta_k\sim q_0} \prod_{i=1}^n (1+\theta_k n^{-1/2} h_{Jk}(Y_i)) \ . \end{aligned} \tag{2.10}$$

Remark that we can not further interchange the ordering of $\mathbf{E}_{\theta_k}$ and $\prod_{i=1}^n$, since the mixture is on data distributions $(Y_1,\dots,Y_n)$ jointly.

Let's introduce the polynomial $f(\theta_k; h_{jk}(Y_1),\dots,h_{jk}(Y_n))$ in θ_k (and $h_{Jk}(Y_i)$) with degree at most n appearing in the above expression, which will be used extensively in the next step,

$$\begin{aligned} f(\theta_k; h_{jk}(Y_1),\dots,h_{jk}(Y_n)) &:= \prod_{i=1}^n (1+\theta_k \frac{h_{Jk}(Y_i)}{\sqrt{n}}) \\ = \sum_{l=0}^n \theta_k^l \frac{\sum_{i_1<\dots<i_l} h_{Jk}(Y_{i_1})\dots h_{Jk}(Y_{i_l})}{n^{l/2}} &=: \sum_{l=0}^n \theta_k^l \frac{H_{Jk}^{(l)}(Y_1,\dots,Y_n)}{n^{l/2}} \ . \end{aligned} \tag{2.11}$$

Here $H_{JK}^{(l)}(Y_1,\dots,Y_n)$ is a sum of monomials of order l, i.e., $\binom{n}{l}$ terms with each of the form $h_{Ji_1}(Y_{i_1})\dots h_{Jk}(Y_{i_l})$. Denote $f^{[\leqslant K]}, f^{[>K]}$ to be the corresponding truncated polynomial according to the degree. In this convenient notation, we know

$$p_0(Y_1,\dots,Y_n) = \prod_{k\in[2^{dJ}]} \mathop{\mathbf{E}}_{\theta_k\sim q_0} f(\theta_k; h_{Jk}(Y_1),\dots,h_{Jk}(Y_n)). \tag{2.12}$$

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

$$\forall \theta_k, \quad \int_{\mathcal{Y}^{\otimes n}} f(\theta_k; h_{Jk}(y_1),\dots,h_{Jk}(y_n)) dy_1\dots dy_n = 1 \ . \tag{2.13}$$

And the following property according to q_0 and q_1 constructed in Proposition 2.1: $\forall y_1,\dots,y_n$

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

Step 4: total variation, telescoping and the sum-product trick. When there is no confusion, we use $f(\theta_k; h_{Jk}(y^{\otimes n}))$ to abbreviate $f(\theta_k; h_{Jk}(y_1),\dots,h_{Jk}(y_n))$. Recall (2.10), we have

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

We claim that the following telescoping Lemma holds. The proof can be seen clearly through writing the left hand side as a telescoping sum and using the triangle inequality.

Proposition 2.2 (Telescoping). *For $N \in \mathbb{N}, N \geqslant 2$, and $a_i, b_i \geqslant 0, 1 \leqslant i \leqslant N$,*

$$|\prod_{k\in[1,N]} a_k - \prod_{k\in[1,N]} b_k| \leqslant \sum_{i\in[1,N]} |a_i - b_i| \cdot \prod_{k\in[1,i)} b_k \cdot \prod_{k\in(i,N]} a_k \ . \tag{2.14}$$

To make use of the above Lemma, define

$$a_k(h_{Jk}(y_1), \ldots, h_{Jk}(y_n)) := \mathop{\mathbf{E}}_{\theta_k\sim q_1} f(\theta_k; h_{Jk}(y^{\otimes n})) \tag{2.15}$$

$$b_k(h_{Jk}(y_1), \ldots, h_{Jk}(y_n)) := \mathop{\mathbf{E}}_{\theta_k\sim q_0} f(\theta_k; h_{Jk}(y^{\otimes n})) \tag{2.16}$$

Using the the above telescoping proposition, we claim

$$\begin{aligned}\mathrm{TV}(p_1, p_0) &\leqslant \sum_{k\in[2^{dJ}]} \int |a_k - b_k| \cdot \left(\prod_{k'\in[0,k)} b_{k'} \prod_{k''\in(k,2^{dJ}-1]} a_{k''} dy_1 \ldots dy_n \right) \\ &= \sum_{k\in[2^{dJ}]} \mathop{\mathbf{E}}_{\substack{\theta_{k'}\sim q_0, k'\in[0,k) \\ \theta_{k''}\sim q_1, k''\in(k,2^{dJ}-1]}} \mathop{\mathbf{E}}_{Y_1,\ldots,Y_n\sim\nu_{\theta_{-k}}} |a_k(h_{Jk}(Y_1), \ldots, h_{Jk}(Y_n)) - b_k(h_{Jk}(Y_1), \ldots, h_{Jk}(Y_n))|.\end{aligned} \tag{2.17}$$

The reasoning behind the last line is as follows. Firstly, we need to define a tilted measure $\nu_{\theta_{-k}}$ without the influence of the k-the coordinate θ_k,

$$\rho_{\nu_{\theta_{-k}}}(x) = \frac{d\nu_{\theta_{-k}}}{dx} := 1 + \frac{1}{\sqrt{n}} \sum_{\substack{k'\neq k \\ 0\leqslant k'\leqslant 2^{dJ}-1}} \theta_{k'} h_{Jk'}(x) = \prod_{\substack{k'\neq k \\ 0\leqslant k'\leqslant 2^{dJ}-1}} \left(1 + \frac{1}{\sqrt{n}}\theta_{k'} h_{Jk'}(x)\right) \ . \tag{2.18}$$

From the properties established in Step 3, one can verify that

$$\begin{aligned}\mathop{\mathbf{E}}_{\substack{\theta_{k'}\sim q_0, k'\in[0,k) \\ \theta_{k''}\sim q_1, k''\in(k,2^{dJ}-1]}} \prod_{i=1}^n \rho_{\nu_{\theta_{-k}}}(y_i) &= \mathop{\mathbf{E}}_{\substack{\theta_{k'}\sim q_0, k'\in[0,k) \\ \theta_{k''}\sim q_1, k''\in(k,2^{dJ}-1]}} \prod_{i=1}^n \prod_{\substack{k'\neq k \\ 0\leqslant k'\leqslant 2^{dJ}-1}} \left(1 + \frac{1}{\sqrt{n}}\theta_{k'} h_{Jk'}(y_i)\right) \\ &= \mathop{\mathbf{E}}_{\substack{\theta_{k'}\sim q_0, k'\in[0,k) \\ \theta_{k''}\sim q_1, k''\in(k,2^{dJ}-1]}} \prod_{\substack{k'\neq k \\ 0\leqslant k'\leqslant 2^{dJ}-1}} \prod_{i=1}^n \left(1 + \frac{1}{\sqrt{n}}\theta_{k'} h_{Jk'}(y_i)\right) \\ &= \prod_{k'\in[0,k)} b_{k'} \prod_{k''\in(k,2^{dJ}-1]} a_{k''} \ .\end{aligned} \tag{2.19}$$

Now we have proved (2.17), since by using Fubini's theorem,

$$\begin{aligned}&\int |a_k - b_k| \cdot \left(\prod_{k'\in[0,k)} b_{k'} \prod_{k''\in(k,2^{dJ}-1]} a_{k''} dy_1 \ldots dy_n \right) \\ &= \mathop{\mathbf{E}}_{\substack{\theta_{k'}\sim q_0, k'\in[0,k) \\ \theta_{k''}\sim q_1, k''\in(k,2^{dJ}-1]}} \int |a_k(h_{Jk}(y_1), \ldots, h_{Jk}(y_n)) - b_k(h_{Jk}(y_1), \ldots, h_{Jk}(y_n))| \prod_{i=1}^n \rho_{\nu_{\theta_{-k}}}(y_i) dy_1 \ldots dy_n \ .\end{aligned}$$

Let's analyze the term

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

where $Y_1,\ldots Y_n$ are i.i.d. sampled from a measure $\nu_{\theta_{-k}}$. We emphasize 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 the wavelet basis, we know that the random variables

$$h_{Jk}(Y_i)$$

are only determined by $\nu_{\theta_{-k}}$ restricted to the domain of h_{Jk}. Equivalently, the distributions of $h_{Jk}(Y)$'s are the same when $Y\sim\nu_{\theta_{-k}}$ and $Y\sim\mu$. Hence for $Y_1,\ldots,Y_n\sim\nu_{\theta_{-k}}$,

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

Now one can directly bound the TV metric between the complex sum-product distribution p_0 and p_1 defined in (2.10),

$$\begin{aligned}2\mathrm{TV}(p_1,p_0) &\leqslant \sum_{k=0}^{2^{dJ}-1} \mathop{\mathbf{E}}_{Y_1,\ldots,Y_n\sim\mu} |a_k(h_{Jk}(Y_1),\ldots,h_{Jk}(Y_n)) - b_k(h_{Jk}(Y_1),\ldots,h_{Jk}(Y_n))|\\ &= \sum_{k=0}^{2^{dJ}-1} \int \left| \mathop{\mathbf{E}}_{\theta_k\sim q_1} f(\theta_k; h_{Jk}(y^{\otimes n})) - \mathop{\mathbf{E}}_{\theta_k\sim q_0} f(\theta_k; h_{Jk}(y^{\otimes n})) \right| dy_1\ldots dy_n. \qquad (2.20)\end{aligned}$$

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 $2K$ (claimed by Proposition 2.1),

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

First, observe the ℓ_2 bound

$$\int |g_1-g_2| d\mu \leqslant \left(\int (g_1-g_2)^2 d\mu\right)^{1/2} \ . \qquad (2.21)$$

Let's bound the ℓ_2 form, which takes the form

$$\begin{aligned}&\int \left(\mathop{\mathbf{E}}_{\theta_k\sim q_1} f(\theta_k; h_{Jk}(y^{\otimes n})) - \mathop{\mathbf{E}}_{\theta_k\sim q_0} f(\theta_k; h_{Jk}(y^{\otimes n})) \right)^2 dy_1\ldots dy_n \qquad (2.22)\\ =& \mathop{\mathbf{E}}_{\theta,\theta'\sim q_1} \int f(\theta; h_{Jk}(y^{\otimes n})) f(\theta'; h_{Jk}(y^{\otimes n})) dy^{\otimes n} + \mathop{\mathbf{E}}_{\omega,\omega'\sim q_0} \int f(\omega; h_{Jk}(y^{\otimes n})) f(\omega'; h_{Jk}(y^{\otimes n})) dy^{\otimes n}\\ &- 2 \mathop{\mathbf{E}}_{\theta\sim q_1,\omega\sim q_0} \int f(\theta; h_{Jk}(y^{\otimes n})) f(\omega; h_{Jk}(y^{\otimes n})) dy^{\otimes n} \ .\end{aligned}$$

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

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

and

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

Therefore we have for (2.22)

$$\begin{aligned}(2.22) &= \mathop{\mathbf{E}}_{\theta,\theta'\sim q_1}\left[\left(1+\theta\theta'\frac{1}{n}\right)^n\right] + \mathop{\mathbf{E}}_{\omega,\omega'\sim q_0}\left[\left(1+\omega\omega'\frac{1}{n}\right)^n\right] - 2\mathop{\mathbf{E}}_{\theta\sim q_1,\omega\sim q_0}\left[\left(1+\theta\omega\frac{1}{n}\right)^n\right] \\ &= \sum_{l=1}^{\lfloor n/2\rfloor}\left(\mathop{\mathbf{E}}_{\theta,\theta'\sim q_1}[(\theta\theta')^{2l}] + \mathop{\mathbf{E}}_{\omega,\omega'\sim q_0}[(\omega\omega')^{2l}] - 2\mathop{\mathbf{E}}_{\theta\sim q_1,\omega\sim q_0}[(\theta\omega)^{2l}]\right)\frac{\binom{n}{2l}}{n^{2l}} \\ &= \sum_{l=1}^{\lfloor n/2\rfloor}\left(\left(\mathop{\mathbf{E}}_{q_1}[\theta^{2l}]\right)^2 + \left(\mathop{\mathbf{E}}_{q_0}[\theta^{2l}]\right)^2 - 2\mathop{\mathbf{E}}_{q_1}[\theta^{2l}]\mathop{\mathbf{E}}_{q_0}[\theta^{2l}]\right)\frac{\binom{n}{2l}}{n^{2l}}\end{aligned}$$

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

$$\mathop{\mathbf{E}}_{\theta\sim q_1}[\theta^{2l}] = \mathop{\mathbf{E}}_{\theta\sim q_0}[\theta^{2l}] \quad\Rightarrow\quad \left(\mathop{\mathbf{E}}_{q_1}[\theta^{2l}]\right)^2 + \left(\mathop{\mathbf{E}}_{q_0}[\theta^{2l}]\right)^2 - 2\mathop{\mathbf{E}}_{q_1}[\theta^{2l}]\mathop{\mathbf{E}}_{q_0}[\theta^{2l}] = 0 \tag{2.23}$$

therefore the above summation equals

$$\begin{aligned}(2.22) &= \sum_{l=K+1}^{\lfloor n/2\rfloor}\left(\left(\mathop{\mathbf{E}}_{q_1}[\theta^{2l}]\right)^2 + \left(\mathop{\mathbf{E}}_{q_0}[\theta^{2l}]\right)^2 - 2\mathop{\mathbf{E}}_{q_1}[\theta^{2l}]\mathop{\mathbf{E}}_{q_0}[\theta^{2l}]\right)\frac{\binom{n}{2l}}{n^{2l}} \\ &\leqslant \sum_{l=K+1}^{\lfloor n/2\rfloor} 4\tau^{4l}\frac{1}{(2l)!} \\ &\lesssim 4\frac{\tau^{4K}}{(2K)!}\exp(\tau^4) .\end{aligned}$$

Assemble the two bounds, we have

$$\int\left|\mathop{\mathbf{E}}_{\theta_k\sim q_1} f(\theta_k; h_{Jk}(y^{\otimes n})) - \mathop{\mathbf{E}}_{\theta_k\sim q_0} f(\theta_k; h_{Jk}(y^{\otimes n}))\right| dy_1\ldots dy_n \tag{2.24}$$

$$\leqslant 2\frac{\tau^{2K}}{\sqrt{(2K)!}}\exp(\tau^4/2) \tag{2.25}$$

Step 6: combine all pieces. Now continuing (2.20), we have

$$\begin{aligned}
2\mathrm{TV}(p_1,p_0) &\leqslant \sum_{k=0}^{2^{dJ}-1} \mathop{\mathbf{E}}_{Y_1,\ldots,Y_n\sim\mu} |a_k(h_{Jk}(Y_1),\ldots,h_{Jk}(Y_n)) - b_k(h_{Jk}(Y_1),\ldots,h_{Jk}(Y_n))| \\
&= \sum_{k=0}^{2^{dJ}-1} \int \left| \mathop{\mathbf{E}}_{\theta_k\sim q_1} f(\theta_k; h_{Jk}(y^{\otimes n})) - \mathop{\mathbf{E}}_{\theta_k\sim q_0} f(\theta_k; h_{Jk}(y^{\otimes n})) \right| dy_1 \ldots dy_n \\
&\leqslant 2^{dJ}\cdot 2\frac{\tau^{2K}}{\sqrt{2K!}}\exp(\tau^4/2) \lesssim \exp(c\log n - K\log K) \ .
\end{aligned}$$

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

$$2\mathrm{TV}(p_1,p_0) \leqslant \exp(-\frac{c}{2}\log n) \leqslant n^{-c/2}. \tag{2.26}$$

By the construction of the composite hypothesis, we have

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

Denote $\mathcal{D}_n$ to be the collection of data $(Y_1,\cdots,Y_n)$, which is drawn from the distribution $Pr(y^{\otimes n}|\theta) := \prod_{i=1}^n \rho_{\nu_\theta}(y_i)$. For any functional of θ, and for any estimator based on n-i.i.d. samples, we know

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

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

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

where the posterior distribution $p_i(y^{\otimes n}) = \int Pr(y^{\otimes n}|\theta)Q_i(d\theta)$, 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,

$$F(\theta) := d_{\mathcal{F}_\gamma}(\mu, \nu_\theta) = (2^{-dJ})^{-\frac{\beta+\gamma}{d}} \left[\frac{1}{2^{dJ}} \sum_{k\in[2^{dJ}]} |\theta_k| \right] ,$$

hence we know

$$\begin{aligned} |\mathop{\mathbf{E}}_{\theta\sim Q_0} F(\theta) - \mathop{\mathbf{E}}_{\theta\sim Q_1} F(\theta)| &= |\mathop{\mathbf{E}}_{\nu_\theta\sim \mathcal{P}_0} d_{\mathcal{F}_\gamma}(\mu, \nu_\theta) - \mathop{\mathbf{E}}_{\nu_\theta\sim \mathcal{P}_1} d_{\mathcal{F}_\gamma}(\mu, \nu_\theta)| \\ &\gtrsim n^{-\frac{\beta+\gamma}{2\beta+d}} \cdot \frac{\log\log(n)}{\log(n)} \\ 1 - d_{TV}(p_0, p_1) &\geqslant 1 - n^{-c/2} \quad \text{by (2.26)} \\ \frac{\delta_{Q_0} + \delta_{Q_1}}{2} &\lesssim n^{-\frac{\beta+\gamma}{2\beta+d}} \frac{1}{\sqrt{2^{dJ}}} \ll n^{-\frac{\beta+\gamma}{2\beta+d}} \cdot \frac{\log\log(n)}{\log(n)} . \end{aligned}$$

Therefore we have

$$\inf_{\widehat{T}_n} \sup_{\nu\in \mathsf{C}^\beta} \mathbf{E}\, |\widehat{T}_n - d_{\mathcal{F}_\gamma}(\mu, \nu)| \gtrsim n^{-\frac{\beta+\gamma}{2\beta+d}} \cdot \frac{\log\log(n)}{\log(n)} . \tag{2.27}$$

2.2 Proof of the Upper Bound

The upper bound can be obtained through similar derivations as in [14; 20; 26]. We include here for completeness.

The estimator is of the plug-in form, with

$$d_{\mathcal{F}_\gamma}(\widetilde{\mu}_m, \widetilde{\nu}_n) := \sup_{f\in\mathcal{F}_\gamma} |\int f d\widetilde{\mu}_m - \int f d\widetilde{\nu}_n| \tag{2.28}$$

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

$$|d_{\mathcal{F}_\gamma}(\widetilde{\mu}_m, \widetilde{\nu}_n) - d_{\mathcal{F}_\gamma}(\mu, \nu)| \leqslant \sup_{f\in\mathcal{F}_\gamma} |\int f d\widetilde{\mu}_m - \int f d\mu| + \sup_{f\in\mathcal{F}_\gamma} |\int f d\widetilde{\nu}_n - \int f d\nu|. \tag{2.29}$$

Now let's bound $\sup_{f\in\mathcal{F}_\gamma} |\int f d\widetilde{\nu}_n - \int f d\nu|$ 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 as

$$\frac{d\widetilde{\nu}_n}{dx} := \sum_{j=0}^{J} \sum_{k=0}^{2^{dj}-1} \widehat{\mathbf{E}}[h_{jk}] h_{jk}(x) . \tag{2.30}$$

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

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

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

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

for $d \geqslant 2\gamma$.

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

$$d\nu/dx \in \mathsf{B}_\infty^{\beta,\infty} \Rightarrow \forall j,k,\ |v_{jk}| \leqslant (2^{-dj})^{\frac{\beta}{d}+\frac{1}{2}} \tag{2.31}$$

$$f \in \mathsf{B}_\infty^{\gamma,\infty} \Rightarrow \forall j,k,\ |f_{jk}| \leqslant (2^{-dj})^{\frac{\gamma}{d}+\frac{1}{2}} \tag{2.32}$$

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

Balancing the two terms, we have

$$\sup_{\nu\in\mathcal{G}_\beta}\mathbf{E}\sup_{f\in\mathcal{F}_\gamma}|\int f d\widetilde{\nu}_n - \int f d\nu| \lesssim \frac{1}{\sqrt{n}}(2^{dJ})^{\frac{1}{2}-\frac{\gamma}{d}} + (2^{dJ})^{-\frac{\beta+\gamma}{d}} \tag{2.33}$$

$$\asymp n^{-\frac{\beta+\gamma}{2\beta+d}}, \quad \text{with } 2^{dJ} \asymp n^{\frac{1}{2\beta/d+1}} \ . \tag{2.34}$$

Put everything together, we know

$$\mathbf{E}\,|d_{\mathcal{F}_\gamma}(\widetilde{\mu}_m,\widetilde{\nu}_n) - d_{\mathcal{F}_\gamma}(\mu,\nu)| \leqslant (n\wedge m)^{-\frac{\beta+\gamma}{2\beta+d}}. \tag{2.35}$$

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[2] 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.

[3] Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[4] T Tony Cai and Mark G Low. Testing composite hypotheses, hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041, 2011.

[5] Albert Cohen. *Numerical analysis of wavelet methods*, volume 32. Elsevier, 2003.

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

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

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

[9] Jianqing Fan. On the estimation of quadratic functionals. *The Annals of Statistics*, 19(3): 1273–1294, 1991.

[10] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[11] Wolfgang Härdle, Gerard Kerkyacharian, Dominique Picard, and Alexander Tsybakov. *Wavelets, approximation, and statistical applications*, volume 129. Springer Science & Business Media, 2012.

[12] Giovanni Leoni. *A first course in Sobolev spaces*. American Mathematical Soc., 2017.

[13] 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.

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

[15] Tengyuan Liang. On the minimax optimality of estimating the wasserstein metric. *arXiv preprint arXiv:1908.10324*, 2019.

[16] Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. *arXiv preprint arXiv:1802.06132*, 2018.

[17] Shuang Liu, Olivier Bousquet, and Kamalika Chaudhuri. Approximation and convergence properties of generative adversarial learning. In *Advances in Neural Information Processing Systems*, pages 5545–5553, 2017.

[18] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. In *Advances in neural information processing systems*, pages 700–709, 2018.

[19] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[20] 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.

[21] Max Sommerfeld and Axel Munk. Inference for empirical wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):219–238, 2018.

[22] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

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

[24] Ananya Uppal, Shashank Singh, and Barnaás Póczos. Nonparametric density estimation under besov ipm losses. *arXiv preprint arXiv:1902.03511*, 2019.

[25] 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.

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