MATH 4995 – Oral Progress Report

Minimax estimation of smooth densities in Wasserstein distance[1]
Jonathan Niles-Weed & Quentin Berthet

SU Zhaohao

The Hong Kong University of Science and Technology

November 20, 2023

Table of Contents

Optimal Transport

Definition (Transport plan)

A transport plan between two probabily measures $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$ is a probability measure $\pi \in \mathcal{P}(X \times Y)$, whose marginals are μ and ν. The space of transport plans is denoted $\Pi(\mu, \nu)$, i.e.,

$$\Pi(\mu, \nu) = \{\pi \in \mathcal{P}(X \times Y) \mid \pi(A \times Y) = \mu(A),\ \pi(X \times B) = \nu(B)\}$$

Definition (Wasserstein space)

Given a metric space (X, d), for $p \geq 1$, denote $\mathcal{P}_p(X)$ by the space of probability measures with finite pth moment on X, Wasserstein space is the case that $\mathcal{P}_p(X)$ equip with the p-Wasserstein distance, which is defined as

$$W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int d(x, y)^p \pi(dxdy) \right)^{1/p}$$

Optimal Transport

Proposition (Weakly convergence and continuity, Theorem 6.9, [2])

For a polish space (X, d) and $p \in [1, \infty)$, if $(\mu_k)_{k \in \mathbb{N}}$ is a sequence of measures in $\mathcal{P}_p(X)$ and μ is another measure in it, then $\mu_k \rightharpoonup \mu$ iff $W_p(\mu_k, \mu) \to 0$. As a consequence, $W_p(\mu_k, \nu_k) \to 0$ given that $\mu_k \rightharpoonup \mu$, $\nu_k \rightharpoonup \nu$.

As W_p describes the discrepancies between distributions robustly and handles the case effectively when the density functions are intractable, it has been widely used as a loss function in optimization problems over measures.

Most of them involve optimizing functionals of the form: $\nu \mapsto W_p(\nu, \mu)$ where μ is the target measure. As μ might be intractable, scientists replace μ with the empirical measure $\hat{\mu}_n$: $\nu \mapsto W_p(\nu, \hat{\mu}_n)$ when given n i.i.d. samples from μ, while this replacement arouses the study about the convergence rates of the difference $W_p(\mu, \hat{\mu}_n)$.

Optimal Transport

This rates are usually of order $n^{-1/d}$ [3], which indicates that it suffers from curse of dimensionality [4], here d is the dimension of Euclidean space where μ locates at. When the support of μ is somewhat low-dimensional, it can be improved [5].

- Consistency: $W_p(\hat{\mu}_n, \mu) \to 0$
- Curse of dimensionality: $\mathbb{E}W_p(\hat{\mu}_n, \mu) = O(n^{-1/d})$
- If μ's support is m-dimensional: $\mathbb{E}W_p(\hat{\mu}_n, \mu) = O(n^{-1/m})$

To beat curse of dimensionality, recall that sufficient smoothness can substantially mitigate it, this paper proposed the usage of **wavelet estimator** in nonparametric density estimation [6], instead of $\hat{\mu}_n$ and firstly established a connection to **Besov norms** of negative smoothness and general Wasserstein distances.

Wavelets and Besov Spaces

Definition (Orthonormal Basis)

A system of functions $\{\varphi_k, k \in \mathbb{Z}\} \subset L_2(\mathbb{R})$ is called orthonormal basis (ONB) of a subspace $V \subset L_2(\mathbb{R})$, if it is an orthonormal system (ONS): $\langle \varphi_i, \varphi_j \rangle_{L_2(\mathbb{R})} = \delta_{ij}$, and any function $f \in V$ has a representation

$$f(x) = \sum_k c_k \varphi_k(x)$$

where the coefficients c_k satisfy $\sum_k |c_k|^2 < \infty$.

Wavelet basis is ONB, but different from the well-known one: trigonometric basis. Trigonometric basis "localizes" the function in the frequency domain only, while the wavelet basis "localize" it both in the frequency domain and time domain. To better understand this, let's construct a wavelet basis by ourselves [7]!

Wavelets and Besov Spaces

1. Choose your father:
   first pick a suitable $\varphi \in L_2(\mathbb{R})$ which is called *father wavelet*, such that $\{\varphi_{0k} = \varphi(x-k), k \in \mathbb{Z}\}$ is an ONS, and define $\varphi_{jk}(x) = 2^{j/2}\varphi(2^j x - k)$ for $j, k \in \mathbb{Z}$, which generate the following linear spaces:

$$
\begin{aligned}
V_0 &= \{f(x) = \sum_k c_k \varphi(x-k) : \sum_k |c_k|^2 < \infty\} \\
V_1 &= \{h(x) = f(2x) : f \in V_0\} \\
&\vdots \\
V_j &= \{h(x) = f(2^j x) : f \in V_0\}
\end{aligned}
$$

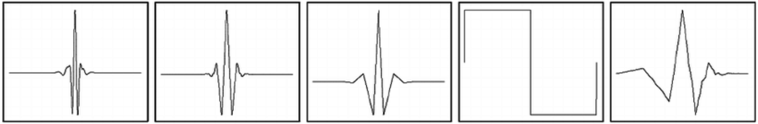
   our φ should be chosen in such a way that $V_j \subset V_{j+1}$ (then $\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots$), and $\cup_j V_j$ is dense in $L_2(\mathbb{R})$. One simple option for φ is that $\varphi(x) = I\{x \in (0,1]\}$ in Haar basis, and V_j consists of the functions in $L_2(\mathbb{R})$ that are constant on the interval of the form $(k2^{-j}, (k+1)2^{-j}]$, $k \in \mathbb{Z}$.

Wavelets and Besov Spaces

2. Find your mother:
   Denote W_j as the orthogonal complement of V_j in V_{j+1}: $V_{j+1} = V_j \oplus W_j$:

$$V_{j+1} = V_j \oplus W_j = \cdots = V_0 \oplus \bigoplus_{i=0}^{j} W_i, \quad L_2(\mathbb{R}) = \mathrm{cl}\left(\cup_{j\geq 0} V_j\right) = V_0 \oplus \bigoplus_{i=0}^{\infty} W_i$$

   and pick suitable $\psi \in W_0$ which called *mother wavelet*, such that $\{\psi_{0k} = \psi(x-k), k \in \mathbb{Z}\}$ is an ONB of W_0. Then, $\{\phi_{jk} = 2^{j/2}\phi(2^j x - k), k \in \mathbb{Z}\}$ is ONB of W_j. The Mother wavelet of Haar basis is $\psi = 1_{[0,1/2]} - 1_{(1/2,1]}$.



(a) coif5 (b) bior2.6 (c) rbior2.6 (d) haar (e) db3

Figure: Some choices of your mother!

Wavelets and Besov Spaces

3. Lovely children:
The children, φ_{0k}, ψ_{jk} are pleased to help their parents to represent any function $f \in L_2(\mathbb{R})$ uniquely:

$$f(x) = \sum_k \alpha_k \varphi_{0k}(x) + \sum_{j=0}^{\infty} \sum_k \beta_{jk} \psi_{jk}(x)$$

where $\alpha_k = \langle f, \varphi_k \rangle_{L_2(\mathbb{R})}$, $\beta_{jk} = \langle f, \psi_{jk} \rangle_{L_2(\mathbb{R})}$. And it can also be done only by girls ψ_{jk}, if we keep decomposing $V_0 = V_{-1} \oplus W_{-1} = \cdots = \oplus_{j<0} W_j$, then we get $f = \sum_j \sum_k \beta_{jk} \psi_{jk}$.

When f is a density function and $X_1, \cdots, X_n$ are i.i.d. samples draw from f, we replace α_k, β_{jk} with $\tilde{\alpha}_k, \tilde{\beta}_{jk}$ in the equation above hence get a **wavelet estimator** $\tilde{f}$ of f, where $\tilde{\alpha}_k, \tilde{\beta}_{jk}$ are defined as

$$(\tilde{\alpha}_k, \tilde{\beta}_{jk}) = \left(\frac{1}{n} \sum_{i=1}^{n} \varphi_{0k}(X_i), \frac{1}{n} \sum_{i=1}^{n} \psi_{jk}(X_i) \right) \rightarrow \left(\int \varphi_{0k} f, \int \psi_{jk} f \right)$$

Wavelets and Besov Spaces

In $\Omega = [0,1]^d \subset \mathbb{R}^d$, assuming the existence of sets $\Phi = \{\phi, \phi \in \Phi\}$ and $\Psi_j = \{\psi, \psi \in \Psi_j\}$ for $j \geq 0$ of functions in $L_2(\Omega)$ satisfying the standard requirements of a wavelet basis, the article used a sequence norm in Besov space:

Definition (Sequence norm & Besov Space, [1])

Suppose $s > 0$ and $p, q \geq 1$, for any $f \in L_p(\Omega)$, $\|f\|_{\mathcal{B}^s_{p,q}}$ is defined as

$$\|f\|_{\mathcal{B}^s_{p,q}} := \|\alpha\|_{l_p} + \left\| 2^{js} 2^{dj(\frac{1}{2}-\frac{1}{p})} \|\beta_j\|_{l_p} \right\|_{l_q}$$

where $\alpha = \{\alpha_\phi\}_{\phi\in\Phi}$ is the vector defined by $\alpha_\phi = \int f\phi$ and $\beta_j = \{\beta_\psi\}_{\psi\in\Psi_j}$ is the vector defined by $\beta_\psi = \int f\psi$, and Besov space $\mathcal{B}^s_{p,q}(\Omega)$ is the set of functions with finite Besov norm, which is equivalent to this sequence norm.

Remark: the index s here measures the smoothness of f, actually $f \in \mathcal{B}^s_{p,q}(\Omega)$ requires that f is $[s]$ times weakly differentiable and $f^{(i)} \in L_p(\mathbb{R}^d)$ for $1 \leq i \leq [s]$.

Wavelets and Besov Spaces

Assumption (Standard requirements, Appendix E in [1])

The constructions of wavelets for $[0,1]^d$ are typically obtained by constructing wavelets for $[0,1]$ and taking products to extend to $[0,1]^d$ by tensorization, then adding suitable functions to guarantee that the resulting sets satisfy the first two assumptions.

1. $\Phi \cup \{\bigcup_{j\geq 0} \Psi_j\}$ is an ONB of $L_2([0,1]^d)$.
2. The functions in Φ and Ψ_j for $j \geq 0$ all lie in $\mathcal{C}^r(\Omega)$, and polynomials of degree at most r on Ω lie in Span(Φ).
3. Each $\psi \in \Psi_j$ is a product of univariate functions ψ_i: $\psi(\mathbf{x}) = \prod_{i=1}^d \psi_i(x_i)$.
4. For each $\psi \in \Psi_j$ there exists a rectangle $I_\psi \subset [0,1]^d$ such that supp$(I_\psi) \subset I_\psi$, diam$(I_\psi) \lesssim 2^{-j}$, and $\left\|\sum_{\psi\in\Psi_j} \mathbf{1}\{x \in I_\psi\}\right\|_{L_\infty} \lesssim 1$.
5. $\|\psi\|_{L_p(\Omega)} \asymp 2^{dj(1/2-1/p)}$ for each $\psi \in \Psi_j$.
6. $\|\nabla f\|_{L_p(\Omega)} \lesssim 2^j \|f\|_{L_p(\Omega)}$ for any $f \in$ Span$(\Phi \cup \{\bigcup_{0\leq k<j} \Psi_k\})$.

Wavelets and Besov Spaces

With these assumptions, we have some straightforward consequences:

1. $|\Phi| \lesssim 1$ and $|\Psi_j| \lesssim 2^{dj}$ for $j \geq 0$ ($|\Psi_j| \asymp 2^{dj}$).
2. For any vector $\{\alpha_\phi\}_{\phi\in\Phi}$ and $\{\beta_\psi\}_{\psi\in\Psi_j}$, then $\left\|\sum_{\phi\in\Phi}\alpha_\phi\phi\right\|_{L_p} \asymp \|\alpha\|_{l_p}$, $\left\|\sum_{\psi\in\Psi_j}\beta_\psi\psi\right\|_{L_p} \asymp 2^{dj(1/2-1/p)}\|\beta\|_{l_p}$.
3. Let P_j denote the orthogonal projection onto the span of Ψ_j, then $\|P_j f\|_{L_p} \lesssim \|f\|_{L_p}$.

Wavelets and Besov Spaces

In this paper, the results mainly focused on the case that μ is compactly supported and bounded from below, hence it is convenient to introduce the following definitions:

Definition ($\mathcal{B}^s_{p,q}(L)$ & $\mathcal{B}^s_{p,q}(L;m)$)

Given $m, L > 0$, set

$$\mathcal{B}^s_{p,q}(L) := \left\{ f \in L_p(\Omega) : \|f - \mathbf{1}\|_{\mathcal{B}^s_{p,q}} \leq L, \int f = 1, f \geq 0 \right\}$$

$$\mathcal{B}^s_{p,q}(L;m) := \mathcal{B}^s_{p,q}(L) \cap \{f : f \geq m\}$$

where $\mathbf{1}$ denotes the constant function taking the value 1 on all Ω.

Table of Contents

Compactly supported & bounded

All the results for the compactly supported and bounded case strongly rely on Theorem 4, which converts the control of Wasserstein distance into that of Besov sequence norm which is more tractable as a norm of functions under the bounded assumptions of $f \vee g$, while the results of $W_p(\hat{\mu}_n, \mu)$ usually don't involve such assumptions.

Theorem (1)

For any $p \geq 1$ and $s > 0$, there exists an estimator $\hat{f}$ such that for any $m > 0$, $p \leq p' < \infty$ and $1 \leq q \leq \infty$, the estimator satisfies

$$\sup_{f \in \mathcal{B}^s_{p',q}(L;m)} \mathbb{E} W_p(f, \hat{f}) \lesssim \begin{cases} n^{-\frac{1+s}{d+2s}}, & d \geq 3 \\ n^{-\frac{1}{2}} \log n, & d = 2 \\ n^{-1/2}, & d = 1 \end{cases}$$

Compactly supported & bounded

The upper bound of Theorem 1 is achieved by a modified wavelet estimator $\hat{f}$:

$$\tilde{f} = \sum_{\phi \in \Phi} \tilde{\alpha}_\phi \phi + \sum_{j=0}^{J} \sum_{\psi \in \Psi_j} \tilde{\beta}_\psi \psi \qquad \hat{f} := \arg\min_{g \in \mathcal{D}} \|\tilde{f} - g\|_{\mathcal{B}^{-1}_{p,1}}$$

it first choose $\tilde{f}$ truncated to a level J which would be specified in the proof.

Note that $\tilde{f}$ might not be a distribution, then define $\hat{f}$ to be the distirbution which is the closest to $\tilde{f}$, and due to this definition we have

$$\|f - \hat{f}\|_{\mathcal{B}^{-1}_{p,1}} \leq \|\tilde{f} - f\|_{\mathcal{B}^{-1}_{p,1}} + \|\tilde{f} - \hat{f}\|_{\mathcal{B}^{-1}_{p,1}} \leq 2\|\tilde{f} - f\|_{\mathcal{B}^{-1}_{p,1}}$$

so we just need to consider $\tilde{f}$, instead of $\hat{f}$ that we even don't know what it would be like!

Compactly supported & bounded

Theorem (2)

For any $p \geq 1$ *and* $L > 0$, *there exists an estimator* $\hat{f}^{\circ}$ *such that for any* $m > 0$, $p \leq p' < \infty$, $1 \leq q \leq \infty$ *and* $s > 0$, *the estimator satisfies*

$$\sup_{f \in \mathcal{B}^s_{p',q}(L;m)} \mathbb{E} W_p(f, \hat{f}^{\circ}) \lesssim \begin{cases} n^{-\frac{1+s}{d+2s}} \log n, & d \geq 3 \\ n^{-\frac{1}{2}} (\log n)^2, & d = 2 \\ n^{-1/2}, & d = 1 \end{cases}$$

It's natural to compare the results of Theorem 2 and that of Theorem 1: $\hat{f}^{\circ}$ in Theorem 2 is adaptive to the smoothness while the one of Theorem 1 is not, but at the price of an extra logarithmic factor and elegance. Though $\hat{f}^{\circ}$ is so ugly that I wouldn't show you here, it shares similar ideas as $\hat{f}$ in Theorem 1.

Compactly supported & bounded

Theorem (3)

For any $p, p', q \geq 1$ *and* $s > 0$,

$$\inf_{\tilde{\mu}} \sup_{f \in \mathcal{B}^s_{p',q}(L;m)} \mathbb{E} W_p(f, \tilde{\mu}) \gtrsim \begin{cases} n^{-\frac{1+s}{d+2s}}, & d \geq 2 \\ n^{-1/2}, & d = 1 \end{cases}$$

where the infimum is taken over all estimators $\tilde{\mu}$ *based on* n *observations.*

The lower bound given by Theorem 3 indicates the tightness of the previous results up to a log factor. The proofs of lower bound are quite different from the one of upper bound and easier, they relate to the minimax theory (now you know what is minimax!) in nonparametric density estimation and usually directly apply **Assouad's Lemma** to draw the conclusions.

Compactly supported & bounded

Theorem (4)

Let $p \in [1, +\infty)$*,* f, g *are two densities in* $L_p([0,1]^d)$*, assume* $M \geq f(x) \vee g(x) \geq m > 0$ *for almost every* $x \in [0,1]^d$*. Then*

$$M^{-1/p'} \|f - g\|_{\mathcal{B}^{-1}_{p,\infty}} \lesssim W_p(f, g) \lesssim m^{-1/p'} \|f - g\|_{\mathcal{B}^{-1}_{p,1}}$$

where $p^{-1} + p'^{-1} = 1$*.*

Here we are, the core technical contribution in this part, or even in this paper! It allows us to bypass the difficulties of W_p by bounding instead a nearly equivalent norm under the bounded assumptions. Prior work only explored the case $p = 2$, and some proofs of similar results are even wrong.

The proof of Theorem 4 is based on **Benamou-Brenier formula**, which interpret W_p in a beautiful fluid dynamic perspective and connects the cost with energy.

Compactly supported & unbounded

The case that the densities are no longer bounded from below is more challenging as we can not use function norms to controal W_p like what we did in Theorem 4. Acutally it can be shown that $\sup_{f,g\in\mathcal{D}} W_p(f,g)/\left\|f-g\right\| = +\infty$ for any given function norm $\left\|\cdot\right\|$ when $p>1$ (as W_1 is still a functional norm).

The methods in this case are closely related to the **dyadic partition**, which are widely used in the estimation of $W_p(\hat{\mu}_n,\mu)$, as we pointed out at the begining, the discrete case doesn't rely on the bounded assumptions.

Theorem (5)

For any $p, p', q \geq 1$ and $s>0$, if L is a sufficiently large constant, then

$$\inf_{\tilde{\mu}} \sup_{f\in\mathcal{B}^s_{p',q}(L)} \mathbb{E}W_p(f,\tilde{\mu}) \gtrsim \begin{cases} n^{-\frac{1+s/p}{d+s}}, & d-s\geq 2p \\ n^{-\frac{1}{2p}}, & d-s<2p \end{cases}$$

where the infimum is taken over all estimators $\tilde{\mu}$ based on n observations.

Compactly supported & unbounded

The results of Theorem 5 are quite different from the ones in previous subsection, as we can see it depends on the dimension p of the Wasserstein distance we choose.

For the case $d - s < 2p$, it's even hard to tell whether it is better than the discrete case $\mathbb{E}W_p(\hat{\mu}_n, \mu)$ of order $n^{-1/d}$.

For the case $d - s \geq 2p$, the smoothness helps but it is still strictly worse than the upper bound given in Theorem 1 when $p \geq 2$ for all $s > 0$ and $d \geq 2p + s$:

$$\frac{1 + s/p}{d + s} \geq \frac{1 + 2s/(d - s)}{d + s} = \frac{1}{d - s} > \frac{1}{d}$$

$$\frac{1 + s/p}{d + s} - \frac{1 + s}{d + 2s} = \frac{s}{(d + s)(d + 2s)} \left[s\left(\frac{2}{p} - 1\right) + \left(\frac{d}{p} + 1 - d\right) \right] < 0$$

Compactly supported & unbounded

Theorem (6)

Assume $p \geq 2$. For any $s \in [0,1)$, there exists a histogram estimator $\hat{f}$ such that

$$\sup_{f \in \mathcal{C}^s(L)} \mathbb{E} W_p(f, \hat{f}) \lesssim \begin{cases} n^{-\frac{1+s/p}{d+s}}, & d-s > 2p \\ n^{-\frac{1}{2p}} \log n, & d-s = 2p \\ n^{-\frac{1}{2p}}, & d-s < 2p \end{cases}$$

$\mathcal{C}^s(L)$ here stands for the Hölder class.

The upper bounds are limited to the case $s < 1$ because they rely on particular properties of the Haar wavelet basis. Denote V_j for the span of $\Phi \cup \{\cup_{0 \leq k < j} \Psi_k\}$ of Haar wavelet, the functions in V_j are precisely those which are constant on the elements of $\mathcal{Q}_j$, which consists of all cubes of the form

$$\mathcal{Q} = [k_1 2^{-j}, (k_1+1)2^{-j}) \times \cdots \times [k_d 2^{-j}, (k_d+1)2^{-j}), \qquad (k_1, \cdots, k_d) \in \mathbb{Z}^d$$

Non-compactly supported & sub-Gaussian

Theorem (7)

Assume $p \geq 2$. For any $s \in [0,1)$, there exists a histogram estimator $\hat{f}$ such that

$$\sup_{f \in \mathcal{C}^s(L;\sigma^2)} \mathbb{E} W_p(f, \hat{f}) \lesssim \begin{cases} n^{-\frac{1+s/p}{d+s}} (\log n)^{\frac{d}{2p}}, & d - s > 2p \\ n^{-\frac{1}{2p}} (\log n)^{1+\frac{d}{2p}}, & d - s = 2p \\ n^{-\frac{1}{2p}} \{\log n \vee (\log n)^{\frac{d}{2p}}, & d - s < 2p \end{cases}$$

$\mathcal{C}^s(L;\sigma^2)$ stands for the set of probability densities on $\mathbb{R}^d$ with s-Hölder norm bounded by L that satisfy $\mathbb{E}_{X \sim f} \exp(\|X\|^2 / 2d\sigma^2) \leq 2$.

Acutally this theorem wasn't included in the arxiv versions of this paper. It is an extentsion of Theorem 6 on the unbounded densities with sub-Gaussian tails, at the price of additional logarithmic factors. It first considers the case with compact support then enlarges the support gradually to the sub-Gaussian case.

Table of Contents

Dynamic formulations

Theorem (1)

For any $p \geq 1$ and $s > 0$, there exists an estimator $\hat{f}$ such that for any $m > 0$, $p \leq p' < \infty$ and $1 \leq q \leq \infty$, the estimator satisfies

$$\sup_{f \in \mathcal{B}^s_{p',q}(L;m)} \mathbb{E} W_p(f, \hat{f}) \lesssim \begin{cases} n^{-\frac{1+s}{d+2s}}, & d \geq 3 \\ n^{-\frac{1}{2}} \log n, & d = 2 \\ n^{-1/2}, & d = 1 \end{cases}$$

Theorem (4)

Let $p \in [1, +\infty)$, f, g are two densities in $L_p([0,1]^d)$, assume $M \geq f(x) \vee g(x) \geq m > 0$ for almost every $x \in [0,1]^d$. Then

$$M^{-1/p'} \|f - g\|_{\mathcal{B}^{-1}_{p,\infty}} \lesssim W_p(f, g) \lesssim m^{-1/p'} \|f - g\|_{\mathcal{B}^{-1}_{p,1}}$$

where $p^{-1} + p'^{-1} = 1$.

Dynamic formulations

The upper bound of Theorem 1 is achieved by a modified wavelet estimator $\hat{f}$:

$$\tilde{f} = \sum_{\phi \in \Phi} \tilde{\alpha}_\phi \phi + \sum_{j=0}^{J} \sum_{\psi \in \Psi_j} \tilde{\beta}_\psi \psi \qquad \hat{f} := \arg\min_{g \in \mathcal{D}} \|\tilde{f} - g\|_{\mathcal{B}_{p,1}^{-1}}$$

it first choose $\tilde{f}$ truncated to a level J which would be specified in the proof.

Note that $\tilde{f}$ might not be a distribution, then define $\hat{f}$ to be the distirbution which is the closest to $\tilde{f}$, and due to this definition we have

$$\|f - \hat{f}\|_{\mathcal{B}_{p,1}^{-1}} \le \|\tilde{f} - f\|_{\mathcal{B}_{p,1}^{-1}} + \|\tilde{f} - \hat{f}\|_{\mathcal{B}_{p,1}^{-1}} \le 2\|\tilde{f} - f\|_{\mathcal{B}_{p,1}^{-1}}$$

so we just need to consider $\tilde{f}$, instead of $\hat{f}$ that we even don't know what it would be like!

Dynamic formulations

Proof of Theorem 1.
Define f_J as the projection of f to Span$(\Phi \cup (\bigcup_{j \le J} \Psi_j))$ (compare it with $\tilde{f}$):

$$f_J = \sum_{\phi \in \Phi} \alpha_\phi \phi + \sum_{j=0}^{J} \sum_{\psi \in \Psi_j} \beta_\psi \psi, \qquad \tilde{f} = \sum_{\phi \in \Phi} \tilde{\alpha}_\phi \phi + \sum_{j=0}^{J} \sum_{\psi \in \Psi_j} \tilde{\beta}_\psi \psi$$

1. f is close to f_J (if J is large):
The assumption that $f \in \mathcal{B}^s_{p',q}(L; m)$ implies that

$$2^{js} 2^{dj(\frac{1}{2}-\frac{1}{p})} \|\beta_j\|_{l_p} \le 2^{js} 2^{dj(\frac{1}{2}-\frac{1}{p'})} \|\beta_j\|_{l_{p'}} \le \|f\|_{\mathcal{B}^s_{p',q}} \lesssim L+1$$

hence we have

$$\|f - f_J\|_{\mathcal{B}^{-1}_{p,1}} = \sum_{j>J} \sum_{\psi \in \Psi_j} 2^{-j} 2^{dj(\frac{1}{2}-\frac{1}{p})} \|\beta_j\|_{l_p} \lesssim \sum_{j>J} \sum_{\psi \in \Psi_j} 2^{-j} \cdot 2^{-js} \asymp 2^{-J(s+1)}$$

Dynamic formulations

2. f_J is close to $\tilde{f}$ (if n is large):
For this case, it is necessary to analyze the errors between estimated coefficients and the real ones, and we introduce the following proposition

Proposition (4)

Let $p \geq 1$ and $0 \leq j \leq J$, if $n \geq 2^{dj}$, then

$$\mathbb{E}\,\|\alpha - \tilde{\alpha}\|_{l_p} \lesssim \frac{1}{\sqrt{n}}, \quad \mathbb{E}\|\beta_j - \tilde{\beta}_j\|_{l_p} \lesssim \frac{2^{dj/p}}{\sqrt{n}}$$

This proposition was derived from **Rosenthal Inequality** [8], which gives an upper bound of the moment of the sum of independent variables: $\mathbb{E}|\sum_i (X_i - \mathbb{E}X_i)|^p$.

$$\begin{aligned}\mathbb{E}\|f_J - \tilde{f}\|_{\mathcal{B}^{-1}_{p,1}} &\lesssim \mathbb{E}\|\alpha - \tilde{\alpha}\|_{l_p} + \sum_{0\leq j\leq J} 2^{-j} 2^{dj(\frac{1}{2}-\frac{1}{p})} \cdot \mathbb{E}\|\beta_j - \tilde{\beta}_j\|_{l_p} \\ &\lesssim n^{-1/2} + \sum_{0\leq j\leq J} 2^{-j} 2^{dj/2} \asymp \sum_{0\leq j\leq J} (2^{d/2-1})^j \cdot n^{-1/2}\end{aligned}$$

Dynamic formulations

3. f is close to $\tilde{f}$ hence close to $\hat{f}$ (if n and J are both large):

$$\begin{aligned} \mathbb{E}W_p(f,\hat{f}) &\lesssim m^{-1/p'}\mathbb{E}\|f-\hat{f}\|_{\mathcal{B}^{-1}_{p,1}} \lesssim \mathbb{E}\|f-\tilde{f}\|_{\mathcal{B}^{-1}_{p,1}} \\ &\le \|f-f_J\|_{\mathcal{B}^{-1}_{p,1}} + \mathbb{E}\|f_J-\tilde{f}\|_{\mathcal{B}^{-1}_{p,1}} \\ &\lesssim 2^{-J(s+1)} + \sum_{0\le j\le J}(2^{d/2-1})^j \cdot n^{-1/2} \end{aligned}$$

this holds as long as $n \ge 2^{dJ}$. Choose J so that $n \asymp 2^{J(d+2s)}$, $2^J \asymp n^{\frac{1}{d+2s}}$.

If $d \ge 3$, the sum and the first term are of the same order as

$$2^{(d/2-1)J} \cdot n^{-1/2} \asymp n^{\frac{d/2-1}{d+2s}-1/2} \asymp n^{-\frac{1+s}{d+2s}} \asymp 2^{J(s+1)}$$

If $d \le 2$, J becomes smaller, hence the sum dominates and is of order $n^{-1/2}$ if $d=1$, or $n^{-1/2}\log n$ if $d=2$.

Dynamic formulations

Proposition (Dynamic perspective, [9])

When $X \subset \mathbb{R}^d$, consider a fluid follow the distribution $\rho_0 = \mu$ at time $t = 0$ and $\rho_1 = \nu$ at time $t = 1$ (and ρ_t for time t). We may wonder how it completes this evolution in such unit time.

For each time t, there is a velocity field $u_t : \mathbb{R}^d \to \mathbb{R}^d$ which stands for the velocity of the flux. The relation between u_t and ρ_t is given by the continuity equation (*Eulerian* description): $\partial_t \rho_t = -\nabla \cdot (\rho_t u_t)$, i.e., for any $\varphi \in C_c^\infty(\mathbb{R}^d)$

$$\frac{d}{dt} \int \varphi(x) \rho_t(dx) = -\int \varphi[\nabla \cdot (\rho_t u_t)] = \int \langle \nabla \varphi, u_t \rangle \, \rho_t(dx)$$

When u_t is regular enough to define its flow map $T : [0,1] \times X \to \mathbb{R}^d$ as the solution of the corresponding PDE: $\dot{T}_t(x) = u_t(T_t(x))$ with $T_t(x) = x$, the flow map here means $T_t(x)$ gives the position at time t of a particle which located at x at time 0, and *Lagrangian* description just tells us: $\rho_t = (T_t)_{\#}\rho_0$.

Dynamic formulations

Eulerian and *Lagrangian*'s descriptions are two ways that describe the same process of flow and are equivalent to some extent. To connect them with optimal transport, we consider the *cost*: the generalized kinetic energy functional

$$A_p(\rho, u) := A_p\left((\rho_t, u_t)_{t\in[0,1]}\right) = \int_0^1 \int \|u_t\|^p \, d\rho_t dt$$

Proposition (Benamou-Brenier formula, [10])

Given two compactly supported measures $\rho_0, \rho_1 \in \mathcal{P}(\mathbb{R}^d)$ and $p \geq 1$, then

$$\begin{aligned} W_p^p(\rho_0, \rho_1) &= \inf\left\{A_p(\rho, u) : (\rho_t, u_t)_{t\in[0,1]} \text{ solves } \partial_t \rho_t = -\mathrm{div}(\rho_t u_t)\right\} \\ &= \inf\left\{\int_0^1 \int \|u_t\|^p \, d\rho_t dt : (\rho_t, u_t)_{t\in[0,1]} \text{ solves } \partial_t \rho_t = -\mathrm{div}(\rho_t u_t)\right\} \end{aligned}$$

Dynamic formulations

Theorem (4)

Let $p \in [1, +\infty)$*,* f, g *are two densities in* $L_p([0,1]^d)$*, assume* $M \geq f(x) \vee g(x) \geq m > 0$ *for almost every* $x \in [0,1]^d$*. Then*

$$M^{-1/p'} \|f - g\|_{\mathcal{B}^{-1}_{p,\infty}} \lesssim W_p(f, g) \lesssim m^{-1/p'} \|f - g\|_{\mathcal{B}^{-1}_{p,1}}$$

where $p^{-1} + p'^{-1} = 1$*.*

Proof.

We first consider the upper bound. Assume there exists such a vector field V on $\Omega = [0,1]^d$ that $V \cdot \mathbf{n} = 0$ on $\partial\Omega$ and $\nabla \cdot V = f - g$, which imply that

$$\int_\Omega h(f - g) dx = \int_\Omega h[\nabla \cdot V] = -\int_\Omega \nabla h \cdot V$$

for any $h \in C^1(\Omega)$. Based on such V, we can construct a flow $(\rho_t, u_t)_{t \in [0,1]}$ that $\partial_t \rho_t = -\mathrm{div}(\rho_t u_t)$: $\rho_t(x) = (1 - \lambda(t)) f(x) + \lambda(t) g(x)$, $u_t(x) = \lambda'(t) V(x) / \rho_t(x)$.

Dynamic formulations

(continued)
The choice of $\lambda(t)$ is tricky, some papers just simply applied $\lambda(t) = t$ to get similar results, which were fallacious [11] [12]. Here we define $\lambda(t) : [0,1] \to [0,1]$ piecewise:

$$\lambda(t) := \begin{cases} (2t)^p/2 & \text{if } t \le 1/2 \\ 1-(2-2t)^p/2 & \text{if } t > 1/2 \end{cases}$$

define $k(t) = \min\{\lambda(t), 1-\lambda(t)\}$, what makes this choice of $\lambda(t)$ special is that

$$\lambda'(t) = \begin{cases} p(2t)^{p-1} & \text{if } t \le 1/2 \\ p(2-2t)^{p-1} & \text{if } t > 1/2 \end{cases} = p(2k(t))^{(p-1)/p}$$

i.e., $(\lambda'(t))^p = 2^{p-1}p^p \cdot k(t)^{p-1}$, which implies

$$\begin{aligned} \rho_t(x) &= (1-\lambda(t))f(x) + \lambda(t)g(x) \ge \max\{(1-\lambda(t))f(x), \lambda(t)g(x)\} \\ &\ge \max\{k(t)f(x), k(t)g(x)\} \ge k(t)m \end{aligned}$$

Dynamic formulations

(continued)
According to Brenier formula and the construction of $(\rho_t, u_t)_{t\in[0,1]}$ here, we know

$$
\begin{aligned}
W_p^p(f,g) &\le \int_0^1 \int_\Omega \|u_t\|^p \, d\rho_t dt = \int_0^1 \int_\Omega \left\| \frac{\lambda'(t)V(x)}{\rho_t(x)} \right\|^p \rho_t(x) dx dt \\
&\le \int_0^1 \int_\Omega \|V(x)\|^p \cdot \frac{(\lambda'(t))^p}{(k(t)m)^{p-1}} dx dt \\
&= p^p \left(\frac{2}{m}\right)^{p-1} \int_\Omega \|V(x)\|^p \, dx \lesssim m^{1-p} \|V\|_{L_p(\Omega)}^p
\end{aligned}
$$

then $W_p(f,g) \lesssim m^{-1/p'} \|V\|_{L_p(\Omega)}$, for any vector field V satisfies the restrictions. It is natural to consider constructing such a vector field, with additional condition: $\|V\|_{L_p(\Omega)} \lesssim \|f-g\|_{\mathcal{B}^{-1}_{p,1}}$. Moreover, it seems the best result actually exists:

$$
\left| \int \psi(f-g) \right| = \left| \int \psi(\nabla \cdot V) \right| = \left| \int \nabla\psi \cdot V \right| \le \|\nabla\psi\|_{L_{p'}} \|V\|_{L_p}
$$

Dynamic formulations

(continued)
taking the supremum over all ψ with $\|\nabla\psi\|_{L_{p'}} \leq 1$, we can conclude that $\|f-g\|_{\dot{H}_p^{-1}} \leq \|V\|_{L_p}$, where $\|\cdot\|_{\dot{H}_p^{-1}}$ stands for the dual Sobolev norm:

Definition (Dual Sobolev norm, [11])

Let $p \geq 1$, for any signed measure ℓ on $\mathbb{R}^d$ with zero total mass, its dual Sobolev norm is defined as

$$\|\ell\|_{\dot{H}_p^{-1}} := \sup\left\{\ell(f) : f \in C_c^\infty,\ \|\nabla f\|_{L_{p'}} \leq 1\right\}$$

where p' is the conjugate of p.

Constructions of vector field V:
We first define vector fields V_ϕ for each $\phi \in \Phi$ and V_ψ for each $\psi \in \Psi_j$, $j \geq 0$ satisfying $\nabla \cdot V_\phi = \phi$ and $\nabla \cdot V_\psi = \psi$, along with appropriate boundary conditions, then the desired vector field V will be obtained as

Dynamic formulations

(continued)

$$V = \sum_{\phi \in \Phi} (\alpha_\phi - \alpha'_\phi) V_\phi + \sum_{j \geq 0} \sum_{\psi \in \Psi_j} (\beta_\psi - \beta'_\psi) V_\psi$$

where the convergence holds in L_p, here (α, β) and (α', β') are the corresponding wavelet coefficients in the expansions of f and g. And the additional condition $\|V\|_{L_p} \lesssim \|f - g\|_{\mathcal{B}^{-1}_{p,1}}$ would be verified.

For the lower bound, we first introduce a beautiful lemma:

Lemma (2, [13])

Let $p \geq 1$, f and g are two densities of absolutely continuous measures ρ_1 and ρ_0, and assume that $f(x) \vee g(x) \leq M$. For all $h \in C^1(\Omega)$

$$\int_\Omega h(f - g) dx \leq M^{1/p'} \|\nabla h\|_{L_{p'}} W_p(f, g)$$

where p' is the conjugate of p.

Dynamic formulations

Proof.
Let ρ_t be the constant speed geodesic between ρ_0 and ρ_1, and u_t be the velocity field such that $(\rho_t, u_t)_t$ satisfies $\partial_t \rho_t = -\text{div}(\rho_t u_t)$, and $A_p(\rho, u) = W_p^p(\rho_0, \rho_1)$. ρ_t is absolutely continuous, and its density is also bounded by M a.e. , then

$$
\begin{aligned}
\int_\Omega h(f-g)dx &= \int_\Omega h d\rho_1 - \int_\Omega h d\rho_0 = \int_0^1 \frac{d}{dt}\left(\int_\Omega h d\rho_t\right) dt \\
&= \int_0^1 \int_\Omega h \partial_t \rho_t dt = \int_0^1 \int_\Omega \nabla h \cdot u_t d\rho_t dt \\
&\leq \left(\int_0^1 \int_\Omega \|\nabla h\|^{p'} d\rho_t dt\right)^{1/p'} \left(\int_0^1 \int_\Omega \|u_t\|^p d\rho_t dt\right)^{1/p} \\
&\leq \left(M \int_0^1 \int_\Omega \|\nabla h\|^{p'} dx dt\right)^{1/p'} A_p(\rho, u)^{1/p} \\
&= M^{1/p'} \|\nabla h\|_{L_{p'}} W_p(f, g)
\end{aligned}
$$

Dynamic formulations

Lemma (?)

Let $p \geq 1$, f and g are two densities of absolutely continuous measures ρ_1 and ρ_0, and assume that $f(x) \vee g(x) \leq M$. For all $h \in C^1(\Omega)$

$$\int_\Omega h(f-g)dx \leq \|\nabla h\|_{L_\infty} W_1(f,g)$$

where p' is the conjugate of p.

Proof.

Let ρ_t be the constant speed geodesic between ρ_0 and ρ_1, and u_t be the velocity field such that $(\rho_t, u_t)_t$ satisfies $\partial_t \rho_t = -\mathrm{div}(\rho_t u_t)$, and $A_1(\rho, u) = W_1(\rho_0, \rho_1)$.

$$\int_0^1 \int_\Omega \nabla h \cdot u_t d\rho_t dt \leq \|\nabla h\|_{L_\infty} \left(\int_0^1 \int_\Omega \|u_t\| \, d\rho_t dt \right) = \|\nabla h\|_{L_\infty} W_1(f,g)$$

In the following discussion, (p, p') defaults to be $(1, +\infty)$.

Dynamic formulations

Fix an index $j \geq 0$, let h be a function of the form

$$h = \sum_{\phi \in \Phi} \kappa_\phi \phi + \sum_{\psi \in \Psi_j} \lambda_\psi \psi$$

for some vectors κ and λ satisfying $\|\kappa\|_{l_{p'}} \leq 1$ and $\|\lambda\|_{l_{p'}} \leq 2^{-j+dj(\frac{1}{2}-\frac{1}{p})}$.

Lemma (3)

If $\|\kappa\|_{l_{p'}} \leq 1$ and $\|\lambda\|_{l_{p'}} \leq 2^{-j+dj(\frac{1}{2}-\frac{1}{p})}$, then $\|\nabla h\|_{L_{p'}(\Omega)} \lesssim 1$

Lemma 2 and Lemma 3 together imply that

$$\begin{aligned} W_p(f,g) &\gtrsim M^{-1/p'} \|\nabla h\|_{L_{p'}(\Omega)}^{-1} \int_\Omega h(f-g)dx \gtrsim M^{-1/p'} \int_\Omega h(f-g)dx \\ &= M^{-1/p'} \left(\sum_{\phi \in \Phi} \kappa_\phi(\alpha_\phi - \alpha'_\phi) + \sum_{\psi \in \Psi_j} \lambda_\psi(\beta_\psi - \beta'_\psi) \right) \end{aligned}$$

Dynamic formulations

Note that

$$\sum_{\phi\in\Phi}\kappa_\phi(\alpha_\phi-\alpha'_\phi)+\sum_{\psi\in\Psi_j}\lambda_\psi(\beta_\psi-\beta'_\psi)=\kappa\cdot(\alpha-\alpha')+\lambda\cdot(\beta_j-\beta'_j)$$

Taking the supremum over all κ and λ satisfying the $l_{p'}$ constraints:

$$\begin{aligned}\sup_{\kappa,\lambda}\left(\kappa\cdot(\alpha-\alpha')+\lambda\cdot(\beta_j-\beta'_j)\right)&=\sup_{\kappa,\lambda}\left(\|\kappa\|_{l_{p'}}\|\alpha-\alpha'\|_{l_p}+\|\lambda\|_{l_{p'}}\left\|\beta_j-\beta'_j\right\|_{l_p}\right)\\&=\|\alpha-\alpha'\|_{l_p}+2^{-j+dj(\frac{1}{2}-\frac{1}{p})}\left\|\beta_j-\beta'_j\right\|_{l_p}\end{aligned}$$

keep taking the supremum over all $j\geq 0$, then the proof is finished:

$$\begin{aligned}&\sup_{j\geq 0}\left(\|\alpha-\alpha'\|_{l_p}+2^{-j}2^{dj(\frac{1}{2}-\frac{1}{p})}\left\|\beta_j-\beta'_j\right\|_{l_p}\right)\\&=\|\alpha-\alpha'\|_{l_p}+\left\|2^{-j}2^{dj(\frac{1}{2}-\frac{1}{p})}\left\|\beta_j-\beta'_j\right\|_{l_p}\right\|_{l_\infty}=\|f-g\|_{\mathcal{B}^{-1}_{p,\infty}}\end{aligned}$$

Dynamic formulations

Proof of Lemma 3.

Recall that the standard assumptions of the wavelet system show that

1. $\|\nabla f\|_{L_p(\Omega)} \lesssim 2^j \|f\|_{L_p(\Omega)}$ for any $f \in \mathrm{Span}(\Phi \cup \{\bigcup_{0\le k<j} \Psi_k\})$.
2. for any vector $\{\alpha_\phi\}_{\phi\in\Phi}$ and $\{\beta_\psi\}_{\psi\in\Psi_j}$, then

$$\left\|\sum_{\phi\in\Phi} \alpha_\phi \phi\right\|_{L_p} \asymp \|\alpha\|_{l_p}, \qquad \left\|\sum_{\psi\in\Psi_j} \beta_\psi \psi\right\|_{L_p} \asymp 2^{dj(1/2-1/p)} \|\beta\|_{l_p}$$

.

the first one implies that

$$\begin{aligned}\|\nabla h\|_{L_{p'}(\Omega)} &\le \left\|\nabla\left(\sum_{\phi\in\Phi} \kappa_\phi \phi\right)\right\|_{L_{p'}(\Omega)} + \left\|\nabla\left(\sum_{\psi\in\Psi_j} \lambda_\psi \psi\right)\right\|_{L_{p'}(\Omega)} \\ &\lesssim \left\|\sum_{\phi\in\Phi} \kappa_\phi \phi\right\|_{L_{p'}(\Omega)} + 2^j \left\|\sum_{\psi\in\Psi_j} \lambda_\psi \psi\right\|_{L_{p'}(\Omega)}\end{aligned}$$

Dynamic formulations

(continued)
and the second one implies

$$RHS \lesssim \|\kappa\|_{l_{p'}} + 2^{j+dj(\frac{1}{2}-\frac{1}{p'})}\|\lambda\|_{l_{p'}} \leq 1 + +2^{j+dj(\frac{1}{2}-\frac{1}{p'})} \cdot 2^{-j+dj(\frac{1}{2}-\frac{1}{p})} = 2$$

Dynamic formulations

Definition of V_ϕ for $\phi \in \Phi$:
Given $x \in \mathbb{R}^d$, we write $x^{(i)}$ for the vector consisting of the first i coordinates of x. For each $1 \leq i \leq d$, define $\phi^{(i)}: \mathbb{R}^i \to \mathbb{R}$ by

$$\phi^{(i)}\left(x^{(i)}\right) = \int_0^1 \cdots \int_0^1 \phi\left(x^{(i)}, t_{i+1}, \ldots, t_d\right) \mathrm{d}t_{i+1} \ldots \mathrm{d}t_d$$

and set $\phi^{(0)} = 0$. We define V_ϕ componentwise as

$$\left(V_\phi\right)_i(x) = \int_0^{x_i} \phi^{(i)}\left(x^{(i-1)}, t_i\right) \mathrm{d}t_i - x_i \phi^{(i-1)}\left(x^{(i-1)}\right) \quad 1 \leq i \leq d$$

Definition of V_ψ for $\psi \in \Psi_j, j \geq 0.$:
We adopt essentially the same construction as above. First, by Assumption E.3, ψ can be written as $\bigotimes_{i=1}^d \psi_i$, where each ψ_i is a univariate function. Assumptions E.1 and E.2 imply that $\int_{[0,1]^d} \psi(x) \mathrm{d}x = 0$, so there exists an index $k \in [d]$ such that $\int_{[0,1]} \psi_k\left(x_k\right) \mathrm{d}x_k = 0$. We set

Dynamic formulations

$$\left(V_\psi\right)_k(x)=\int_0^{x_k} \psi_k(t) \mathrm{d} t \cdot \prod_{i \neq k} \psi_i\left(x_i\right)$$

and $\left(V_\psi\right)_i=0$ for $i \neq k$.

Dynamic formulations

Dyadic partition

Definition (Dyadic partition, [5])

A dyadic partition of a set $S \subset X$ with parameter $\delta < 1$ is a sequence $\{\mathcal{Q}_k\}_{k=1}^{k_1}$ with $\mathcal{Q}_k \subset \mathcal{B}(X)$ possessing the following properties:

- The sets in $\mathcal{Q}_k$ form a partition of S.
- If $Q \in \mathcal{Q}_k$, then $\operatorname{diam}(Q) \le \delta^k$.
- If $Q_{k+1} \in \mathcal{Q}_{k+1}$ and $Q_k \in \mathcal{Q}_k$, then either $Q_{k+1} \subset Q_k$ or $Q_{k+1} \cap Q_k = \emptyset$.

Proposition ([5])

μ and ν are two probability measures on X. If $\{\mathcal{Q}_k\}_{k=1}^{k_1}$ is a dyadic partition of S with parameter δ, where S is a bounded set S with $\mu(S) = \nu(S) = 1$, then

$$W_p^p(\mu,\nu) \le (\operatorname{diam}(S)^p \vee 1) \cdot \left(\delta^{k_1 p} + \sum_{k=1}^{k_1} \delta^{(k-1)p} \sum_{Q \in \mathcal{Q}_k} |\mu(Q) - \nu(Q)| \right)$$

Dyadic partition

Proof.

For the process from α to β, given a partition $\mathcal{Q}$ we first ensure that their mass on each $Q \in \mathcal{Q}$ equals (or are close): $\alpha(S) \to \beta(S)$, then as the partition $\mathcal{Q}$ is refined enough: $\max_{Q\in\mathcal{Q}} \text{diam}(Q) \to 0$, we deem that α, β are close enough.

To begin with, for $\mathcal{Q}_1$ we scale μ, ν on each $Q_1 \in \mathcal{Q}_1$ to achieve the same mass:

$$\mu_1 = \sum_{Q_1 \in \mathcal{Q}_1} \frac{\mu(Q_1) \wedge \nu(Q_1)}{\mu(Q_1)} \cdot \mu|_{Q_1}, \qquad \nu_1 = \sum_{Q_1 \in \mathcal{Q}_1} \frac{\mu(Q_1) \wedge \nu(Q_1)}{\nu(Q_1)} \cdot \nu|_{Q_1}$$

we have $\mu_1(Q_1) = \nu_1(Q_1) = \mu(Q_1) \wedge \nu(Q_1)$ and $\mu_1 \leq \mu$, $\nu_1 \leq \nu$.

Define $(\pi_1, \rho_1) = (\mu - \mu_1, \nu - \nu_1)$ as the measures that are throwed away during the process $(\mu, \nu) \to (\mu_1, \nu_1)$, and it's easy to check

$$\pi_1(S) = \rho_1(S) = \sum_{Q_1 \in \mathcal{Q}_1} [\mu(Q_1) - \mu_1(Q_1)] = \frac{1}{2} \sum_{Q_1 \in \mathcal{Q}_1} |\mu(Q_1) - \nu(Q_1)|$$

Dyadic partition

(continued)
As $(\mu, \nu) = (\mu_1 + \pi_1, \nu_1 + \rho_1)$ and $W_p^p(\pi_1, \rho_1) \leq (\mathrm{diam}(S))^p \pi_1(S)$, we have

$$W_p^p(\mu, \nu) \leq W_p^p(\pi_1, \rho_1) + W_p^p(\mu_1, \nu_1) \leq \mathrm{diam}(S)^p \pi_1(S) + W_p^p(\mu_1, \nu_1)$$

To continue this refinement, we define $\{\mu_k\}_{k=1}^{k_1}$, $\{\nu_k\}_{k=1}^{k_1}$ by induction

$$\mu_k = \sum_{Q_k \in \mathcal{Q}_k} \frac{\mu_{k-1}(Q_k) \wedge \nu_{k-1}(Q_k)}{\mu_{k-1}(Q_k)} \cdot \mu_{k-1}|_{Q_k}$$

$$\nu_k = \sum_{Q_k \in \mathcal{Q}_k} \frac{\mu_{k-1}(Q_k) \wedge \nu_{k-1}(Q_k)}{\nu_{k-1}(Q_k)} \cdot \nu_{k-1}|_{Q_k}$$

and similarly, $\mu_k(Q_k) = \nu_k(Q_k) = \mu_{k-1}(Q_k) \wedge \nu_{k-1}(Q_k)$ for any $Q_k \in \mathcal{Q}_k$. Let $(\pi_k, \rho_k) = (\mu_{k-1} - \mu_k, \nu_{k-1} - \nu_k)$, the idea could be illustrated as

$$W_p^p(\mu, \nu) = W_p^p\left(\mu_{k_1} + \sum_{k=1}^{k_1} \pi_k, \nu_{k_1} + \sum_{k=1}^{k_1} \rho_k\right) \leq W_p^p(\mu_{k_1}, \nu_{k_1}) + \sum_{k=1}^{k_1} W_p^p(\pi_k, \rho_k)$$

Dyadic partition

(continued)

To estimate $W_p^p(\pi_k, \rho_k)$ when $k > 1$, we first check that

$$\pi_k(S) = \rho_k(S) = 2^{-1} \sum_{Q_k \in \mathcal{Q}_k} |\mu_{k-1}(Q_k) - \nu_{k-1}(Q_k)|$$

and they have the same mass on each $Q_{k-1} \in \mathcal{Q}_{k-1}$:

$$\pi_k(Q_{k-1}) = \mu_{k-1}(Q_{k-1}) - \mu_k(Q_{k-1}) = \nu_{k-1}(Q_{k-1}) - \nu_k(Q_{k-1}) = \rho_k(Q_{k-1})$$

then it's natural to restrict the transport $\pi_k \to \rho_k$ within each $Q \in \mathcal{Q}_{k-1}$:

$$W_p^p(\pi_k, \rho_k) \leq \sum_{Q \in \mathcal{Q}_{k-1}} W_p^p(\pi_k|Q, \rho_k|Q) \leq \sum_{Q \in \mathcal{Q}_{k-1}} \mathrm{diam}(Q)^p \pi_k(Q) \leq \delta^{(k-1)p} \pi_k(S)$$

for $W_p^p(\mu_{k_1}, \nu_{k_1})$ we just give a rough estimation: $W_p^p(\mu_{k_1}, \nu_{k_1}) \leq \delta^{k_1 p}$. Finally the last step, the estimation of $\pi_k(S)$, the sum of $|\mu_{k-1}(Q_k) - \nu_{k-1}(Q_k)|$:

Dyadic partition

(continued)

Recall that on each $Q \in \mathcal{Q}_{k-1}$, μ_{k-1} is proportional to μ, so as ρ_{k-1}, we assume

$$(\mu_{k-1}|Q, \nu_{k-1}|Q) = (a_Q \mu|Q, b_Q \nu|Q), \quad (a_Q, b_Q) = \left(\frac{\mu_{k-1}(Q)}{\mu(Q)}, \frac{\nu_{k-1}(Q)}{\nu(Q)}\right) \in [0,1]$$

note that $\mu_{k-1}(Q) = \nu_{k-1}(Q)$, for any $P \in \mathcal{Q}_k, P \subset Q$ we have:

$$\begin{aligned}
&\sum_{P \subset Q} |\mu_{k-1}(P) - \nu_{k-1}(P)| = \sum_{P \subset Q} |a_Q \mu(P) - b_Q \nu(P)| \\
\leq &\sum_{P \subset Q} [a_Q \cdot |\mu(P) - \nu(P)| + |a_Q - b_Q| \cdot \nu(P)] \\
= &a_Q \sum_{P \subset Q} |\mu(P) - \nu(P)| + |a_Q - b_Q| \cdot \nu(Q) \\
= &a_Q \sum_{P \subset Q} |\mu(P) - \nu(P)| + \frac{\mu_{k-1}(Q)}{\mu(Q)} |\mu(Q) - \nu(Q)| \leq 2 \sum_{P \subset Q} |\mu(P) - \nu(P)|
\end{aligned}$$

Dyadic partition

(continued)
with this we are able to complete the proof:

$$\pi_k(S) = \frac{1}{2}\sum_{P\in\mathcal{Q}_k} |\mu_{k-1}(P) - \nu_{k-1}(P)| = \frac{1}{2}\sum_{Q\in\mathcal{Q}_{k-1}}\sum_{P\subset Q} |\mu_{k-1}(P) - \nu_{k-1}(P)|$$
$$\leq \sum_{Q\in\mathcal{Q}_{k-1}}\sum_{P\subset Q} |\mu(P) - \nu(P)| = \sum_{P\in\mathcal{Q}_k} |\mu(P) - \nu(P)|$$

hence

$$W_p^p(\mu,\nu) \leq \delta^{k_1 p} + \mathrm{diam}(S)^p \pi_1(S) + \sum_{k=2}^{k_1} \delta^{(k-1)p} \sum_{P\in\mathcal{Q}_k} |\mu(P) - \nu(P)|$$
$$\leq (\mathrm{diam}(S)^p \vee 1)\cdot\left(\delta^{k_1 p} + \sum_{k=1}^{k_1} \delta^{(k-1)p} \sum_{Q\in\mathcal{Q}_k} |\mu(Q) - \nu(Q)|\right)$$

Dyadic partition

We should see the strong connection between Haar basis and dyadic partion:

Denote V_j for the span of $\Phi \cup \{\cup_{0\le k<j}\Psi_k\}$ of Haar wavelet, the functions in V_j are precisely those which are constant on the elements of $\mathcal{Q}_j$, which consists of all cubes of the form in $\Omega = [0,1]^d$

$$\mathcal{Q} = [k_1 2^{-j}, (k_1+1)2^{-j}) \times \cdots \times [k_d 2^{-j}, (k_d+1)2^{-j}), \qquad (k_1, \cdots, k_d) \in \mathbb{Z}^d$$

these cubes actually form a dyadic partion with $\delta = 1/2$, denote K_j as the orthogonal projection onto V_j, we have diam$(\Omega) \lesssim 1$ and

$$W_p(\mu,\nu) \lesssim \left(\sum_{j\ge 0} 2^{-jp} \sum_{Q\in\mathcal{Q}_j} |\mu(Q)-\nu(Q)|\right)^{1/p} \lesssim \left(\sum_{j\ge 0} 2^{-jp} \|K_j(f-g)\|\right)^{1/p}$$

$$\lesssim \sum_{j\ge 0} 2^{-jp} \|K_j(f-g)\|^{1/p} = \sum_{j\ge 0} 2^{-jp} \|K_j(f-g)\|_{L_1(\Omega)}^{1/p}$$

To sharpen this, we replace RHS with an L_p bound when f is "steady" on Q.

Dyadic partition

Theorem (11)

Let $J \geq 0$ be an integer, f and g are probability densities on $\mathbb{R}^d$ with support in $[-B, B]^d$ for some integer B and that satisfy $g/f \in V_J$. Then

$$W_p(f,g) \lesssim B \left\| K_0(f-g) \right\|_{L_1(\mathbb{R}^d)}^{1/p} + \sum_{0 \leq j < J} 2^{-j} \Bigg(\sum_{Q \in Q_j} \left\| K_{j+1}(f-g) \right\|_{L_1(Q)} \wedge m_Q^{1-p} \Delta_Q^p \left\| K_{j+1}(f-g) \right\|_{L_p(Q)}^p \Bigg)^{1/p}$$

where

$$m_Q := \frac{\int_Q g}{\int_Q f} \cdot \inf_{x \in Q} f(x) \qquad \Delta_Q := \frac{\sup_{x \in Q} f(x)}{\inf_{x \in Q} f(x)}$$

What could $g/f \in V_J$ imply? What is the benefit to introduce such strange terms $m_Q^{1-p} \Delta_Q^p \left\| K_{j+1}(f-g) \right\|_{L_p(Q)}^p$? When should we use the latter one instead?

Dyadic partition

Proof.
Recall that in the proof of the proposition, we introduce μ_k that is proportional to μ on each $Q \in \mathcal{Q}_k$, then

$$W_p^p(\mu, \nu) = W_p^p\left(\mu_{k_1} + \sum_{k=1}^{k_1} \pi_k, \nu_{k_1} + \sum_{k=1}^{k_1} \rho_k\right) \leq W_p^p(\mu_{k_1}, \nu_{k_1}) + \sum_{k=1}^{k_1} W_p^p(\pi_k, \rho_k)$$

here we follow the same logic: define f_j by setting it proportional to f partially

$$f_j(x) = \frac{\int_Q g}{\int_Q f} f(x) \quad \forall x \in Q, Q \in \mathcal{Q}_j,$$

The assumption $g/f \in V_j$ actually means $f_J = g$, hence

$$W_p(f, g) \leq W_p\left(f, f_0\right) + \sum_{0 \leq j < J} W_p\left(f_j, f_{j+1}\right)$$

We estimate $W_p(f, f_0)$ and $W_p(f_j, f_{j+1})$ separtely.

Dyadic partition

1. $W_p(f, f_0)$:

Lemma ([2])

Let μ and ν be two probability measures on a polish space (X, d), then for any $p \in [1, +\infty)$ and $x_0 \in X$

$$W_p^p(\mu, \nu) \le 2^{p-1} \left(\int_X d(x_0, x)^p d|\mu - \nu|(x) \right)$$

This upper bound can be done from the definition by taking

$$\pi(dxdy) := (\mu \wedge \nu)(dx)\mathbf{1}(x = y) + a^{-1}(\mu - \nu)_+(dx) \otimes (\mu - \nu)_-(dy)$$

as a transport plan with $a = (\mu - \nu)_-(X) = (\mu - \nu)_+(X)$. Here we roughly have

$$\begin{aligned} W_p^p(f, f_0) &\lesssim \text{diam}([-B, B]^d)^p \int |f(x) - f_0(x)| dx \\ &\lesssim (B\sqrt{d})^p \sum_{Q \in \mathcal{Q}_0} \int_Q |f(x) - f_0(x)| dx \end{aligned}$$

Dyadic partition

The definition of f_0 implies for all $Q \in \mathcal{Q}_0$

$$\int_Q |f(x) - f_0(x)|\,\mathrm{d}x = \int_Q \left| \frac{\int_Q f - \int_Q g}{\int_Q f} \right| f(x)\mathrm{d}x = \left| \int_Q f - \int_Q g \right|$$

$$W_p(f, f_0) \lesssim B\sqrt{d} \left(\sum_{Q \in \mathcal{Q}_0} \left| \int_Q f - \int_Q g \right| \right)^{1/p} = B\sqrt{d}\, \|K_0(f-g)\|_{L_1(\mathbb{R}^d)}^{1/p}$$

Dyadic partition

2. $W_p(f_j, f_{j+1})$:

Lemma (C.1)

For $j \geq 0$, $p \geq 1$, if f_j, f_{j+1} are two densities such that $K_j f_j = K_j f_{j+1}$. Then

$$W_p^p(f_j, f_{j+1}) \lesssim 2^{-jp} \sum_{Q \in \mathcal{Q}_j} \|f_j - f_{j+1}\|_{L_1(Q)} \wedge \left(\inf_{x \in Q} f_j(x)\right)^{1-p} \|f_j - f_{j+1}\|_{L_p(Q)}^p$$

Lemma (C.2)

For any $Q \in \mathcal{Q}_j$ and $p \geq 1$, the densities f_j and f_{j+1} satisfy

$$\|f_j - f_{j+1}\|_{L_1(Q)} \lesssim \|K_{j+1}(f-g)\|_{L_1(Q)}$$

$$\|f_j - f_{j+1}\|_{L_p(Q)} \lesssim \Delta_Q \|K_{j+1}(f-g)\|_{L_p(Q)}$$

where $\Delta_Q := (\sup_{x \in Q} f(x))/(\inf_{x \in Q} f(x))$.

Combine all these results finish the proof of Theorem 11.

Dyadic partition

Proof of Lemma C.1

$K_j f_j = K_j f_{j+1}$ implies that f_j and f_{j+1} assign the same mass to each $Q \in \mathcal{Q}_j$, then consider a coupling between f_j and f_{j+1} by coupling $f_j|_Q$ and $f_{j+1}|_Q$:

$$W_p^p\left(f_j, f_{j+1}\right) \leq \sum_{Q \in \mathcal{Q}_j} W_p^p\left(\left.f_j\right|_Q, \left.f_{j+1}\right|_Q\right)$$

For the L_1 bound, the lemma indicates: $W_p^p(\mu, \nu) \lesssim \operatorname{diam}(Q)^p \|\mu - \nu\|_{L_1(Q)}$
For the L_p bound, denote P_k as the L_2 projection onto Span(Ψ_k), $m_Q(f_j)$ as the infimum of f_j on Q, then apply Theorem 4 with the standard assumptions:

$$\begin{aligned} W_p\left(\left.f_j\right|_Q, \left.f_{j+1}\right|_Q\right) &\lesssim m_Q(f_j)^{-1/p'} \left\| f_j|_Q - f_{j+1}|_Q \right\|_{\mathcal{B}_{p,1}^{-1}} \\ &\lesssim m_Q(f_j)^{-1/p'} \sum_{k \geq j} 2^{-k} \left\| P_k\left(f_j - f_{j+1}\right) \right\|_{L_p(Q)} \\ &\lesssim 2^{-j} m_Q(f_j)^{-1/p'} \left\| f_j - f_{j+1} \right\|_{L_p(Q)} \end{aligned}$$

Dyadic partition

Proof of Lemma C.2

Denote μ, ν as the measures with density f, g. Recall that in the proof of proposition, we have $\sum_P |\mu_k(P) - \nu_k(P)| \leq 2 \sum_P |\mu(P) - \nu(P)|$, here we replace μ_k with K_j, and instead we know: the densities $K_{j+1}f$ and $K_{j+1}g$ are constant with values $2^{d(j+1)}\mu(R)$ and $2^{d(j+1)}\nu(R)$ on each $R \in \mathcal{Q}_{j+1}$, assume $R \subseteq Q \in \mathcal{Q}_j$, we have that $(f_j - f_{j+1}) = \left(\frac{\nu(Q)}{\mu(Q)} - \frac{\nu(R)}{\mu(R)}\right) f$ on R, which implies

$$\|f_j - f_{j+1}\|_{L_k(Q)}^k = \sum_{\substack{R \in \mathcal{Q}_{j+1} \\ R \subseteq Q}} \|f\|_{L_k(R)}^k \left|\frac{\nu(Q)}{\mu(Q)} - \frac{\nu(R)}{\mu(R)}\right|^k .$$

When $k = 1$, we know $\|f\|_{L_1(R)} = \mu(R)$ and similarly get

$$\sum_{R \subset Q} \|f\|_{L_1(R)} \left|\frac{\nu(Q)}{\mu(Q)} - \frac{\nu(R)}{\mu(R)}\right| \leq 2 \sum_{R \subset Q} |\mu(R) - \nu(R)| \leq 2 \|K_{i+1}(f - g)\|_{L_1(Q)}$$

Dyadic partition

When $k = p$, since $(a+b)^p \le 2^{p-1}(a^p + b^p)$ we know

$$\left|\frac{\nu(Q)}{\mu(Q)} - \frac{\nu(R)}{\mu(R)}\right|^p \lesssim \frac{1}{\mu(R)^p}|\nu(R) - \mu(R)|^p + \frac{1}{\mu(Q)^p}|\nu(Q) - \mu(Q)|^p$$

Since R is a cube of side length $2^{-(j+1)}$, we have the simple bounds

$$\mu(Q)^p \ge \mu(R)^p \ge 2^{-(j+1)pd}\left(\inf_Q f\right)^p, \qquad \|f\|^p_{L_p(R)} \le 2^{-(j+1)d}\left(\sup_Q f\right)^p$$

hence obtain

$$\|f\|^p_{L_p(R)}\left|\frac{\nu(Q)}{\mu(Q)} - \frac{\nu(R)}{\mu(R)}\right|^p \lesssim \Delta_Q^p 2^{(j+1)(p-1)d}\left(|\nu(R) - \mu(R)|^p + |\nu(Q) - \mu(Q)|^p\right)$$

and applying the definition of $K_{j+1}f$ and $K_{j+1}g$ yields the claim

$$\|f_j - f_{j+1}\|^p_{L_p(Q)} = \sum_{\substack{R\in\mathcal{Q}_{j+1}\\ R\subseteq Q}} \|f\|^p_{L_p(R)}\left|\frac{\nu(Q)}{\mu(Q)} - \frac{\nu(R)}{\mu(R)}\right|^p \lesssim \Delta_Q^p\,\|K_{j+1}(f-g)\|^p_{L_p(Q)}$$

Dyadic partition

Theorem (6)

Assume $p \geq 2$. *For any* $s \in [0, 1)$, *there exists a histogram estimator* $\hat{f}$ *such that*

$$\sup_{f \in \mathcal{C}^s(L)} \mathbb{E} W_p(f, \hat{f}) \lesssim \begin{cases} n^{-\frac{1+s/p}{d+s}}, & d - s > 2p \\ n^{-\frac{1}{2p}} \log n, & d - s = 2p \\ n^{-\frac{1}{2p}}, & d - s < 2p \end{cases}$$

$\mathcal{C}^s(L)$ *here stands for the Hölder class.*

Proof.

Choose Haar wavelet and define our estimator $\hat{f}$ as (where J satisfies $n \asymp 2^{J(d+s)}$)

$$\hat{f} := 1 + \sum_{0 \leq j < J} \sum_{\psi \in \Psi_j} \hat{\beta}_\psi \psi$$

μ, ν are measures of $f, \hat{f}$, then $n\nu(Q) \sim B(n, \mu(Q))$ for $Q \in \mathcal{Q}_j$ with $j < J$.

Dyadic partition

To exploit Theorem 11, we also define a g such that $g/f \in V_J$:

$$g(x) = \frac{\nu(Q)}{\mu(Q)} f(x) \quad \forall x \in Q, Q \in \mathcal{Q}_J$$

Fix $Q \in \mathcal{Q}_j$ with $j < J$, Theorem 11 says the contribution to $W_p(f, g)$ from Q can be bounded by either $\|K_{j+1}(f - g)\|_{L_1(Q)}$ or $m_Q^{1-p} \Delta_Q^p \|K_{j+1}(f - g)\|_{L_p(Q)}^p$. We can always apply the first bound, but we will apply the second when the density f is bounded below on Q.

We denote by $\mathcal{E}_Q$ the high-probability event that $\nu(Q) \geq \frac{1}{2}\mu(Q)$ and divide into cases according to $\mathcal{E}_Q$ and $\inf_Q f$. Actually only the first case dominates, the latter two are asymptotically negligible.

Dyadic partition

1. $\inf_{x\in Q} f(x) < 2^{-js}$:
Under the assumption that $f \in \mathcal{C}^s(L)$, we have

$$\sup_{x\in Q} f(x) \leq \inf_{x\in Q} f(x) + L\,\mathrm{diam}(Q)^s \lesssim 2^{-js}$$

Therefore $\mu(Q) = \int_Q f(x)\mathrm{d}x \lesssim 2^{-j(d+s)}$. As $K_{j+1}f$ and $K_{j+1}g$ are constant on each $R \subseteq Q, R \in \mathcal{Q}_{j+1}$ with value $2^{d(j+1)}\mu(R)$ and $2^{d(j+1)}\nu(R)$. This implies

$$\|K_{j+1}(f-g)\|_{L_1(Q)} = \sum_{R\subseteq Q} 2^{d(j+1)} \int_R |\mu(R) - \nu(R)| = \sum_{R\subseteq Q} |\mu(R) - \nu(R)|$$

In particular, recalling that $n \cdot \nu(R) \sim \mathrm{Bin}(n, \mu(R))$, we have that

$$\mathbb{E}\,\|K_{j+1}(f-g)\|_{L_1(Q)}^2 \lesssim \sum_{R\subseteq Q} \mathbb{E}(\mu(R) - \nu(R))^2 \leq \mu(Q)/n$$

$$\mathbb{E}\,\|K_{j+1}(f-g)\|_{L_1(Q)} \leq (\mu(Q)/n)^{1/2} \lesssim 2^{-j(d+s)/2} n^{-1/2}$$

Dyadic partition

2. $\inf_{x\in Q} f(x) \geq 2^{-js}$ and $\mathcal{E}_Q$ holds:

$$\mathbb{E}\left[m_Q^{1-p}\Delta_Q^p \left\|K_{j+1}(f-g)\right\|_{L_p(Q)}^p \mathbf{1}_{\mathcal{E}_Q}\right] \lesssim 2^{jd\left(\frac{p}{2}-1\right)}\left(\inf_{x\in Q} f(x)\right)^{1-\frac{p}{2}} n^{-p/2}$$
$$\leq 2^{j(d+s)\left(\frac{p}{2}-1\right)} n^{-p/2}$$

3. $\mathcal{E}_Q$ does not hold:

$$\mathbb{E}\left[\left\|K_{j+1}(f-g)\right\|_{L_1(Q)} \mathbf{1}_{\mathcal{E}_Q^C}\right] \lesssim n^{-1}$$

Under the assumption that $n \geq 2^{j(d+s)}$, the first term dominates. Since there are 2^{jd} elements of $\mathcal{Q}_j$ which lie in $[0,1]^d$, we can get the results

$$\mathbb{E}\sum_{Q\in\mathcal{Q}_j} \left\|K_{j+1}(f-g)\right\|_{L_1(Q)} \wedge m_Q^{1-p}\Delta_Q^p \left\|K_{j+1}(f-g)\right\|_{L_p(Q)}^p \lesssim 2^{j(d-s)/2} n^{-1/2}$$

Dyadic partition

Sum them together

$$\mathbb{E}W_p(f,g) \lesssim \sum_{0\le j<J} 2^{-j}2^{j(d-s)/2p}n^{-1/2p} \lesssim \begin{cases} n^{-\frac{1+s/p}{d+s}} & \text{if } d-s>2p \\ n^{-1/2p}\log n & \text{if } d-s=2p \\ n^{-1/2p} & \text{if } d-s<2p \end{cases}$$

On the other hand, we can show that $\mathbb{E}W_p(g,\hat{f}) \lesssim 2^{-J(1+s/p)} \asymp n^{-\frac{1+s/p}{d+s}}$. Since $W_p(f,\hat{f}) \le W_p(f,g) + W_p(g,\hat{f})$, combining these two bounds yields the claim.

Lemma (C.5)

For g and $\hat{f}$ defined as in the proof of Theorem 6, we have

$$\mathbb{E}W_p(g,\hat{f}) \lesssim 2^{-J(1+s/p)}$$

Proof.

By construction, $\hat{f}$ is an element of V_J while g is defined so that $K_J g = K_J\hat{f}$. Combining these two facts yields that $\hat{f} = K_J g$.

Dyadic partition

Moreover, we have for any $Q \in \mathcal{Q}_J$,

$$\|g - \hat{f}\|_{L_1(Q)} = \|g - K_J g\|_{L_1(Q)} = \frac{\nu(Q)}{\mu(Q)} \|f - K_J f\|_{L_1(Q)}$$

Since $\mathbb{E}\nu(Q) = \mu(Q)$, we obtain $\mathbb{E}\|g - \hat{f}\|_{L_1(\Omega)} = \|f - K_J f\|_{L_1(\Omega)}$. Note that $K_J g = K_J \hat{f}$, Lemma C.1 implies

$$\begin{aligned} W_p(g, \hat{f}) &\lesssim 2^{-J} \left(\sum_{Q \in \mathcal{Q}_j} \|g - \hat{f}\|_{L_1(Q)} \right)^{1/p} = 2^{-J} \|g - \hat{f}\|_{L_1(\Omega)}^{1/p} \\ &= 2^{-J} \|g - K_J g\|_{L_1(\Omega)}^{1/p} \end{aligned}$$

Jensen yields $\mathbb{E}W_p(g, \hat{f}) \lesssim 2^{-J} \|f - K_J f\|_{L_1(\Omega)}^{1/p}$. To conclude, we use the fact that if $f \in \mathcal{C}^s(L)$, then for any $x \in Q \in \mathcal{Q}_J$

$$f(x) - K_J f(x) = f(x) - \frac{1}{\mathrm{vol}(Q)} \int_Q f(y) \mathrm{d}y = \frac{1}{\mathrm{vol}(Q)} \int_Q f(x) - f(y) \mathrm{d}y \lesssim \mathrm{diam}(Q)^s$$

Dyadic partition

Therefore $\|f - K_J f\|_{L_\infty(\Omega)} \lesssim 2^{-Js}$, and we obtain

$$\mathbb{E}W_p(g, \hat{f}) \lesssim 2^{-J(1+s/p)}$$

as desired.

Mass transfer

Theorem (7)

Assume $p \geq 2$. For any $s \in [0,1)$, there exists a histogram estimator $\hat{f}$ such that

$$\sup_{f \in \mathcal{C}^s(L;\sigma^2)} \mathbb{E} W_p(f, \hat{f}) \lesssim \begin{cases} n^{-\frac{1+s/p}{d+s}} (\log n)^{\frac{d}{2p}}, & d - s > 2p \\ n^{-\frac{1}{2p}} (\log n)^{1+\frac{d}{2p}}, & d - s = 2p \\ n^{-\frac{1}{2p}} \{\log n \vee (\log n)^{\frac{d}{2p}}, & d - s < 2p \end{cases}$$

$\mathcal{C}^s(L;\sigma^2)$ stands for the set of probability densities on $\mathbb{R}^d$ with s-Hölder norm bounded by L that satisfy $\mathbb{E}_{X\sim f} \exp(\|X\|^2 / 2d\sigma^2) \leq 2$.

Proof.

First try to reduce it to the compactly supported case, given positive integer B, we restrict the density on $[-B, B]^d$ by mass transfer: define a linear operator $T_B : L_1(\mathbb{R}^d) \to L_1(\mathbb{R}^d)$ by setting for any $h \in L_1(R^d)$:

Mass transfer

$$(T_B h)(x) := \begin{cases} 0 & \text{if } x \notin [-B, B]^d \\ h(x) & \text{if } x \in [-B, B]^d \setminus [0,1]^d \\ h(x) + \int_{\mathbb{R}^d \setminus [-B,B]^d} h(x) & \text{if } x \in [0,1]^d \end{cases}$$

it's clear that T_B transfers the mass that lies out of $[-B, B]^d$ uniformly to $[0,1]^d$. If we have n i.i.d. samples $X_1, \cdots, X_n$ drawed from f, then we can also get n i.i.d. samples $X_1', \cdots, X_n'$ drawed from $T_B f$, by setting $X_i' = X_i$ if $X_i \in [B, B]^d$, and $X_i' \sim \text{Unif}([0,1]^d)$ independent of all other random variables if $X_i \notin [B, B]^d$.

Now we use the same skill as on f and $T_B f$ using $\mathbf{X} = (X_1, \cdots, X_n)$ and $\mathbf{X}' = (X_1', \cdots, X_n')$: fix a truncation level J then let

$$\hat{f} := \sum_{\phi \in \Phi} \hat{\alpha}_\phi(\mathbf{X})\phi + \sum_{0 \le j < J} \sum_{\psi \in \Psi_j} \hat{\beta}_\psi(\mathbf{X})\psi, \quad \hat{f}' := \sum_{\phi \in \Phi} \hat{\alpha}_\phi(\mathbf{X}')\phi + \sum_{0 \le j < J} \sum_{\psi \in \Psi_j} \hat{\beta}_\psi(\mathbf{X}')\psi$$

Mass transfer

We claim that as long as B is large, f and $T_B f$, $\hat{f}$ and $\hat{f}'$ would be very close:

Lemma (C.6)

Let $f \in \mathcal{C}^s\left(L; \sigma^2\right)$. For any $C' > 0$, there exists a constant C such that if $B \geq \sqrt{C\sigma^2 \log n}$, then

$$W_p\left(T_B f, f\right) \lesssim n^{-C'}, \qquad \mathbb{E}W_p(\hat{f}', \hat{f}) \lesssim n^{-C'}$$

Now we've already reduce the case on compactly supported: $\hat{f} \to \hat{f}' \to T_B f \to f$, for the remaining part $\mathbb{E}W_p(\hat{f}', T_B f)$, we also introduce an intermedia g:

$$g(x) = \sum_{Q \in \mathcal{Q}_J} 1_Q(x) \cdot \frac{\nu(Q)}{\mu(Q)} \cdot (T_B f)(x)$$

where μ and ν are the probability measures with densities $T_B f$ and $\hat{f}'$. Note that $g/(T_B f) \in V_J$, we can apply Theorem 11, and we also have $n\nu(Q) \sim B(n, \mu(Q))$ for any $Q \in \mathcal{Q}_j$ if $j < J$, hence we have the similar results in Theorem 6's proof:

Mass transfer

$$\mathbb{E}\left\{\left\|K_{j+1}\left(T_B f-g\right)\right\|_{L_1(Q)} \wedge m_Q^{1-p} \Delta_Q^p \left\|K_{j+1}(f-g)\right\|_{L_p(Q)}^p\right\} \lesssim 2^{-j(d+s)/2} n^{-1/2}$$

for $j<J$ and $Q \in \mathcal{Q}_j$ contained inside $[-B,B]^d$, and as there are $(2B)^d 2^{jd}$ elements of $\mathcal{Q}_j$ in $[-B,B]^d$, then

$$\begin{aligned}
\mathbb{E} W_p(T_B f, g) \lesssim\ & B \mathbb{E}\left\|K_0(f-g)\right\|_{L_1(\mathbb{R}^d)}^{1/p} + \sum_{0 \leq j<J} 2^{-j} \mathbb{E}\Bigg(\sum_{Q \in Q_j}\left\|K_{j+1}(f-g)\right\|_{L_1(Q)} \\
& \wedge m_Q^{1-p} \Delta_Q^p \left\|K_{j+1}(f-g)\right\|_{L_p(Q)}^p\Bigg)^{1/p} \\
\lesssim\ & B \mathbb{E}\left\|K_0(T_B f-g)\right\|_{L_1(\mathbb{R}^d)}^{1/p} + \sum_{0 \leq j<J} 2^{-j}\left(B^d 2^{j(d-s)/2} n^{-1/2}\right)^{1/p}
\end{aligned}$$

We also have the following simple bound, which is obtained directly from the definitions of $T_B f$ and g.

Mass transfer

Lemma (C.7)

$$\mathbb{E}\left\|K_0\left(T_B f-g\right)\right\|_{L_1\left(\mathbb{R}^d\right)} \lesssim B^{d/2} n^{-1/2}$$

Applying Theorem 11 in combination with the above estimates yields

$$\mathbb{E} W_p\left(T_B f, g\right) \lesssim B^{d/2p+1} n^{-1/2p}+\sum_{0 \leq j<J} 2^{-j} B^{d/p} 2^{j(d-s)/2p} n^{-1/2p}$$

Since $\max\{d/2p+1, d/p\} \leq \max\{2, d/p\}$, we can replace the first term by $B^2 n^{-1/2p}$ at the price of a constant factor. We obtain

$$\mathbb{E} W_p\left(T_B f, g\right) \lesssim \begin{cases}B^{d/p} n^{-\frac{1+s/p}{d+s}} & d-s>2p \\ B^{d/p} n^{-1/2p} \log n & d-s=2p \\ B^{\max\{2, d/p\}} n^{-1/2p} & d-s<2p\end{cases}$$

As we already finished the part of $W_p(T_B f, g)$, we still need to estimate $W_p(g, \hat{f}')$ to connect $T_B f$ with $\hat{f}'$:

Mass transfer

Lemma (C.8)

$$\mathbb{E}W_p(g,\hat{f}') \lesssim B^{d/p}2^{-J(1+s/p)} = B^{d/p}n^{-\frac{1+s/p}{d+s}}$$

Combining all these results and choosing $B \asymp \sqrt{\log n}$ yields the claim.

Proof of Lemma C.6.

Applying the definition of T_B yields

$$\begin{aligned} W_p^p\left(T_B f, f\right) \leq & 2^{p-1}\int \|x\|^p \left|T_B f(x) - f(x)\right| \mathrm{d}x \\ \lesssim & \int_{\mathbb{R}^d\backslash[-B,B]^d} \left(\|x\|^p + 1\right) f(x)\mathrm{d}x \end{aligned}$$

If $B \geq p\sigma \vee 1$, then on the set $\mathbb{R}^d\backslash[-B,B]^d$, we know $\|x\|^2 \geq dB^2 \geq pd\sigma^2$, then function should be monotonically increasing of $\|x\|^2$:

$$\|x\|^p \cdot \exp(-\|x\|^2/2d\sigma^2) = \left(\|x\|^2 \cdot \exp(-\|x\|^2/pd\sigma^2)\right)^{p/2}$$

Mass transfer

hence

$$\begin{aligned} \|x\|^p + 1 \lesssim \|x\|^p &= \frac{\|x\|^p}{\exp(\|x\|^2/2d\sigma^2)} \cdot \exp\left(\frac{\|x\|^2}{2d\sigma^2}\right) \\ &\leq \frac{d^{p/2}B^p}{\exp(B^2/2\sigma^2)} \cdot \exp\left(\frac{\|x\|^2}{2d\sigma^2}\right) \\ &\lesssim B^p e^{-B^2/2\sigma^2} e^{\|x\|^2/2d\sigma^2} \end{aligned}$$

If f is σ^2-subgaussian, then

$$\begin{aligned} \int_{\mathbb{R}^d \setminus [-B,B]^d} (\|x\|^p + 1) f(x)\mathrm{d}x &\lesssim B^p e^{-B^2/2\sigma^2} \int e^{\|x\|^2/2d\sigma^2} f(x)\mathrm{d}x \\ &\lesssim B^p e^{-B^2/2\sigma^2} \end{aligned}$$

Therefore, if B is a positive integer such that $B \geq \sqrt{C\sigma^2 \log n}$ for a sufficiently large constant C, then this quantity is at most $n^{-C'}$, as desired.

Mass transfer

The second claim follows by an analogous argument:

$$\begin{aligned} W_p^p(\hat{f}, \hat{f}') &\leq 2^{p-1} \int \|x\|^p |\hat{f}(x) - \hat{f}'(x)| \mathrm{d}x \\ &\lesssim \int_{\mathbb{R}^d \setminus [-B,B]^d} \left(\|x\|^p + 1\right) \hat{f}(x) \mathrm{d}x \\ &\lesssim \int_{\mathbb{R}^d \setminus [-B,B]^d} \left(\|x\|^p + 1\right) \mathrm{d}\hat{\mu}(x) \end{aligned}$$

This follows from the fact that $\hat{\mu}(Q) = \int_Q \hat{f}$ for all $Q \in \mathcal{Q}_0$, and each of these elements of $\mathcal{Q}_0$ has constant diameter. Take the expectation:

$$\mathbb{E} W_p^p(\hat{f}, \hat{f}') \leq \mathbb{E} \int_{\mathbb{R}^d \setminus [-B,B]^d} \left(\|x\|^p + 1\right) \mathrm{d}\hat{\mu}(x) = \int_{\mathbb{R}^d \setminus [-B,B]^d} \left(\|x\|^p + 1\right) f(x) \mathrm{d}x$$

while we have already shown to be at most $n^{-C'}$ when $B \geq \sqrt{C\sigma^2 \log n}$ for C sufficiently large. This proves the claim.

Mass transfer

Proof of Lemma C.7.

Since $n \cdot \nu(Q)$ has distribution $\mathrm{Bin}(\mu(Q), n)$ for each $Q \in \mathcal{Q}_0$, we have

$$\mathbb{E}\left\|K_0\left(T_B f-g\right)\right\|_{L_1\left(\mathbb{R}^d\right)}=\sum_{Q \in \mathcal{Q}_0} \mathbb{E}|\mu(Q)-\nu(Q)| \leq \sum_{Q \in \mathcal{Q}_0} \sqrt{\mu(Q) / n}$$

Since there are at most $(2B)^d$ elements of $\mathcal{Q}_0$ for which $\mu(Q) \neq 0$, the Cauchy-Schwartz inequality implies

$$\sum_{Q \in \mathcal{Q}_0} \sqrt{\mu(Q) / n} \leq (2B)^{d/2}\left(\sum_{Q \in \mathcal{Q}_0} \mu(Q) / n\right)^{1/2}=(2B)^{d/2} n^{-1/2}$$

Mass transfer

Proof oF Lemma C.8.

The proof is the same as that of Lemma C.5. We have constructed $\hat{f}' \in V_J$ and g such that $K_J g = \hat{f}'$. As in the proof of Lemma C.5, this implies

$$\mathbb{E}\left\|g - \hat{f}'\right\|_{L_1(\mathbb{R}^d)} = \|T_B f - K_J T_B f\|_{L_1(\mathbb{R}^d)}$$

and thus

$$\mathbb{E}W_p\left(g, \hat{f}'\right) \lesssim 2^{-J} \|T_B f - K_J T_B f\|_{L_1(\mathbb{R}^d)}^{1/p}$$

Since the restriction of $T_B f$ to elements of $\mathcal{Q}_J$ is s-Hölder smooth, we have again that $\|T_B f - K_J T_B f\|_{L_\infty(\mathbb{R}^d)} \lesssim 2^{-Js}$. Finally, since the support of $T_B f$ and $K_J T_B f$ lies in $[-B, B]^d$, we obtain

$$\mathbb{E}W_p\left(g, \hat{f}'\right) \lesssim B^{d/p} 2^{-J(1+s/p)}$$

as claimed.

Minimax lower bounds

Proposition (Assouad's lemma, [6])

Let $\Omega = \{0,1\}^m$ be the set of all binary sequences of length m. $\{\mathbb{P}_\omega, \omega \in \Omega\}$ is a set of 2^m probability measures on $(X, \mathcal{A})$ and denote $\mathbb{E}_\omega$ as the corresponding expectations. Then

$$\inf_{\hat{\omega}} \max_{\omega \in \Omega} \mathbb{E}_\omega \rho(\omega, \hat{\omega}) \geq \frac{m}{2} \inf_{(\omega,\omega'):\rho(\omega,\omega')=1} \inf_{\psi} \Big(\mathbb{P}_\omega(\psi = 0) + \mathbb{P}_{\omega'}(\psi = 1) \Big)$$

where $\rho(\omega, \omega') := |\{i : \omega_i \neq \omega'_i\}|$ is the Hamming distance, $\inf_{\hat{\omega}}$ denotes the infimum over all estimators $\hat{\omega}$ taking values in Ω and where $\inf_\psi$ denotes the infimum over all tests ψ taking values in $\{0,1\}$.

Remark: The infimum over ψ seems to relate to the *total variation* $\mathrm{TV}(\cdot,\cdot)$:

$$\inf_{\psi}[\mathbb{P}_\omega(\psi \neq 0) + \mathbb{P}_{\omega'}(\psi \neq 1)] = \inf_{A}[\mathbb{P}_\omega(A) + \mathbb{P}_{\omega'}(A^c)]$$

$$= 1 - \sup_{A}[\mathbb{P}_{\omega'}(A) - \mathbb{P}_\omega(A)] = 1 - \mathrm{TV}(\mathbb{P}_\omega, \mathbb{P}_{\omega'})$$

Minimax lower bounds

Proof.

Define $\omega = (\omega_1, \cdots, \omega_m)$, so as $\hat{\omega}$. Then

$$\max_{\omega\in\Omega} \mathbb{E}_\omega \rho(\omega, \hat{\omega}) \geq 2^{-m} \sum_{\omega\in\Omega} \mathbb{E}_\omega \rho(\omega, \hat{\omega}) = 2^{-m} \sum_{\omega\in\Omega} \sum_{i=1}^{m} \mathbb{E}_\omega |\omega_i - \hat{\omega}_i|$$
$$= 2^{-m} \sum_{i=1}^{m} \left(\sum_{\omega:\omega_i=1} + \sum_{\omega:\omega_i=0} \right) \mathbb{E}_\omega |\omega_i - \hat{\omega}_i|$$

WLOG, we consider the case $i = 1$

$$\left(\sum_{\omega:\omega_1=1} + \sum_{\omega:\omega_1=0} \right) \mathbb{E}_\omega |\omega_i - \hat{\omega}_i| = \sum_{(\omega_2,\cdots,\omega_m)} [\mathbb{E}_{(1,\cdots,\omega_m)}(1 - \hat{\omega}_1) + \mathbb{E}_{(0,\cdots,\omega_m)} \hat{\omega}_1]$$

then $\hat{\omega}_1$ is no longer important, we can eliminate it with taking the infimum over all $\psi \in \{0, 1\}$:

Minimax lower bounds

(continued)

$$\mathbb{E}_{(1,\cdots,\omega_m)}(1-\hat{\omega}_1)+\mathbb{E}_{(0,\cdots,\omega_m)}\hat{\omega}_1 \geq \inf_{(\omega,\hat{\omega}):\rho(\omega,\hat{\omega})=1}[\mathbb{E}_\omega(1-\hat{\omega}_1)+\mathbb{E}_{\omega'}\hat{\omega}_1]$$

$$\geq \inf_{(\omega,\hat{\omega}):|\omega-\hat{\omega}|=1}\inf_\psi[\mathbb{E}_\omega(1-\psi)+\mathbb{E}_{\omega'}\psi] = \inf_{(\omega,\hat{\omega}):|\omega-\hat{\omega}|=1}\inf_\psi[\mathbb{P}_\omega(\psi=0)+\mathbb{P}_{\omega'}(\psi=1)]$$

hence we have

$$\begin{aligned}
\max_{\omega\in\Omega}\mathbb{E}_\omega\rho(\omega,\hat{\omega}) &\geq 2^{-m}\sum_{i=1}^m\left(\sum_{\omega:\omega_i=1}+\sum_{\omega:\omega_i=0}\right)\mathbb{E}_\omega|\omega_i-\hat{\omega}_i| \\
&\geq 2^{-m}\sum_{i=1}^m 2^{m-1}\inf_{(\omega,\hat{\omega}):|\omega-\hat{\omega}|=1}\inf_\psi[\mathbb{P}_\omega(\psi=0)+\mathbb{P}_{\omega'}(\psi=1)] \\
&= \frac{m}{2}\inf_{(\omega,\omega'):\rho(\omega,\omega')=1}\inf_\psi\Big(\mathbb{P}_\omega(\psi=0)+\mathbb{P}_{\omega'}(\psi=1)\Big)
\end{aligned}$$

Minimax lower bounds

When dealing with the minimax lower bounds of all estimators $\hat{\omega}$ based on n i.i.d. samples, we have $p_\omega(\mathbf{x}) = \otimes_{i=1}^n p_\omega(x_i)$, while the calculation of total variation here is thorny, we use *Hellinger distance* instead:

Definition (Hellinger distance)

P and Q are two probability measures on a measure space X that are absolutely continuous with respect to a measure λ, their Hellinger distance is defined as

$$H^2(P,Q) := \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \lambda(dx) = 2 - 2\int \sqrt{p(x)q(x)}\lambda(dx)$$

where $(p,q) = (dP/d\lambda, dQ/d\lambda)$ are the Radon-Nikodym derivatives.

Hellinger distance handles product measures easily: if $(P,Q) = (\otimes_{i=1}^n P_i, \otimes_{i=1}^n P_i)$

$$1 - \frac{H^2(P,Q)}{2} = \int \sqrt{dPdQ} = \prod_{i=1}^n \left(\int \sqrt{dP_i dQ_i}\right) = \prod_{i=1}^n \left(1 - \frac{H^2(P_i,Q_i)}{2}\right)$$

Minimax lower bounds

We first build connection between Hellinger distance and the total variation:

Lemma

$$\mathsf{TV}(P,Q) \leq \sqrt{1-\left(1-\frac{H^2(P,Q)}{2}\right)^2} \leq 1-\frac{1}{2}\left(1-\frac{H^2(P,Q)}{2}\right)^2$$

Proof.

$$\begin{aligned}\left(1-\frac{H^2(P,Q)}{2}\right)^2 &= \left(\int \sqrt{dPdQ}\right)^2 \leq \left(\int (dP \vee dQ)\right)\left(\int (dP \wedge dQ)\right) \\ &= \left(2-\int (dP \wedge dQ)\right)\left(\int (dP \wedge dQ)\right) \\ &= (1+\mathsf{TV}(P,Q))(1-\mathsf{TV}(P,Q))\end{aligned}$$

Minimax lower bounds

With the relationship above, we can establish a practical conclusion:

Proposition ([14] [15])

Let $\Omega = \{0,1\}^m$, $\{\mu_\omega, \omega \in \Omega\} \subset \mathcal{C}$ has 2^m probability measures on $(X, \mathcal{A})$, where $\mathcal{C}$ is a larger family of probability measures. If d is a metric on $\mathcal{P}(X)$ and

(i). For all $\omega, \omega' \in \Omega$, there exists such $\alpha > 0$ that: $d(\mu_\omega, \mu_{\omega'}) \geq \alpha \cdot \rho(\omega, \omega')$.

(ii). For all $\omega, \omega' \in \Omega$ with $\rho(\omega, \omega') = 1$, there exists such $\beta > 0$ that:

$$H^2(\mu_\omega, \mu_{\omega'}) = \int (\sqrt{d\mu_\omega} - \sqrt{d\mu_{\omega'}})^2 \leq \beta$$

then for all $n \geq 1$ we have (m, α and β might depend on n)

$$\inf_{\hat{\mu}_n} \sup_{\mu \in \mathcal{C}} \mathbb{E}_{\mathbf{X} \sim \otimes_{i=1}^n \mu} \, d(\mu, \hat{\mu}_n(\mathbf{X})) \geq \frac{\alpha m}{8} \left(1 - \frac{\beta}{2}\right)^{2n}$$

Minimax lower bounds

Proof.

Let $\hat{\omega}_n = \arg\min_{\omega\in\Omega} d(\mu_\omega, \hat{\mu}_n) = \hat{\omega}_n(\hat{\mu}_n)$, we have

$$d(\mu_\omega, \mu_{\hat{\omega}_n}) \le d(\mu_\omega, \hat{\mu}_n) + d(\mu_{\hat{\omega}_n}, \hat{\mu}_n) \le 2d(\mu_\omega, \hat{\mu}_n)$$

which helps us convert $\inf_{\hat{\mu}_n}$ into $\inf_{\hat{\omega}_n}$:

$$\begin{aligned}
\inf_{\hat{\mu}_n} \sup_{\mu\in\mathcal{C}} \mathbb{E}d(\mu, \hat{\mu}_n) &\ge \inf_{\hat{\mu}_n} \max_{\omega\in\Omega} \mathbb{E}d(\mu_\omega, \hat{\mu}_n) \ge 2^{-1} \inf_{\hat{\mu}_n} \max_{\omega\in\Omega} \mathbb{E}d(\mu_\omega, \mu_{\hat{\omega}_n(\hat{\mu}_n)}) \\
&\ge 2^{-1} \inf_{\hat{\omega}_n} \max_{\omega\in\Omega} \mathbb{E}d(\mu_\omega, \mu_{\hat{\omega}_n}) \ge (\alpha/2) \inf_{\hat{\omega}_n} \max_{\omega\in\Omega} \mathbb{E}_\omega \rho(\omega, \hat{\omega}_n) \\
&\ge \frac{\alpha m}{4} \inf_{(\omega,\omega'):\rho(\omega,\omega')=1} [1 - \mathsf{TV}(\otimes_{i=1}^n \mu_\omega, \otimes_{i=1}^n \mu_{\omega'})] \\
&\ge \frac{\alpha m}{8} \inf_{(\omega,\omega'):\rho(\omega,\omega')=1} \left(1 - H^2(\otimes_{i=1}^n \mu_\omega, \otimes_{i=1}^n \mu_{\omega'})/2\right)^2 \\
&= \frac{\alpha m}{8} \left[\inf_{(\omega,\omega'):\rho(\omega,\omega')=1} \left(1 - H^2(\mu_\omega, \mu_{\omega'})/2\right)^2\right]^{2n} \ge RHS
\end{aligned}$$

Applications of Assouad's Lemma

Theorem (3)

For any $p, p', q \geq 1$ *and* $s > 0$,

$$\inf_{\tilde{\mu}} \sup_{f \in \mathcal{B}^s_{p',q}(L;m)} \mathbb{E} W_p(f, \tilde{\mu}) \gtrsim \begin{cases} n^{-\frac{1+s}{d+2s}}, & d \geq 2 \\ n^{-1/2}, & d = 1 \end{cases}$$

where the infimum is taken over all estimators $\tilde{\mu}$ *based on* n *observations.*

Proof.

By the monotonicity of the Wasserstein-p distances, it suffices to prove the lower bound for the case $p = 1$. Consider an index J to be specified later and a vector $\varepsilon \in \{\pm 1\}^{|\Psi_J|}$, define:

$$f_\varepsilon = 1 + \frac{1}{4\sqrt{n}} \sum_{\psi \in \Psi_J} \varepsilon_\psi \psi$$

it is easy to check $f_\varepsilon \in \mathcal{B}^s_{p',q}(L;m)$ and bounded, Theorem 4 implies

$$W_1(f_\varepsilon, f_{\varepsilon'}) \gtrsim \|f_\varepsilon - f_{\varepsilon'}\|_{\mathcal{B}^{-1}_{1,\infty}} \asymp 2^{-J} 2^{-dJ/2} / \sqrt{n} \cdot \rho(\varepsilon, \varepsilon')$$

Applications of Assouad's Lemma

(continued)

When $\rho(\varepsilon, \varepsilon') = 1$, we have

$$\int (\sqrt{f_\varepsilon} - \sqrt{f_{\varepsilon'}})^2 = \int \frac{(f_\varepsilon - f_{\varepsilon'})^2}{(\sqrt{f_\varepsilon} + \sqrt{f_{\varepsilon'}})^2} \lesssim \int (f_\varepsilon - f_{\varepsilon'})^2 \lesssim n^{-1}$$

with these two conditions, we apply Assouad's Lemma:

$$\begin{aligned} \inf_{\tilde{\mu}} \sup_{f \in \mathcal{B}^s_{p',q}(L;m)} \mathbb{E} W_p(f, \tilde{\mu}) &\geq \inf_{\tilde{\mu}} \sup_{f \in \mathcal{B}^s_{p',q}(L;m)} \mathbb{E} W_1(f, \tilde{\mu}) \\ &\gtrsim \frac{2^{-J} 2^{-dJ/2}}{\sqrt{n}} \cdot |\Psi_J| \gtrsim \frac{2^{-J} 2^{dJ/2}}{\sqrt{n}} \end{aligned}$$

here choosing J such that $2^J \asymp n^{1/(d+2s)}$ when $d \geq 2$ and $J = 0$ when $d = 1$ yields the claim.

Applications of Assouad's Lemma

Before starting the proof of Theorem 5, we first introduce a straightforward result of Wasserstein distance:

Lemma (5)

Let μ and ν be probability measures on $\mathbb{R}^d$, S and T are two compact set with $d(S,T) \geq c$ and $\mu(S \cup T) = \nu(S \cup T) = 1$. Then $W_p(\mu,\nu) \geq c|\mu(S) - \nu(S)|^{1/p}$.

Proof.

WLOG, assume $\mu(S) > \nu(S)$, we derive from the definition of Wasserstein distance:

$$\begin{aligned} W_p^p(\mu,\nu) &= \inf_{\substack{X\sim\mu \\ Y\sim\nu}} \mathbb{E}|X-Y|^p \geq \inf_{\substack{X\sim\mu \\ Y\sim\nu}} \mathbb{E}(\mathbf{1}\{X\in S, Y\in T\}|X-Y|^p) \\ &\geq c^p \inf_{\substack{X\sim\mu \\ Y\sim\nu}} \mathbb{P}(X\in S, Y\in T) \geq c^p \inf_{\substack{X\sim\mu \\ Y\sim\nu}} [\mathbb{P}(X\in S) - \mathbb{P}(Y\notin T)] \\ &\geq c^p|\mu(S)-\nu(S)| \end{aligned}$$

With this lemma, we can now concerntrate on the choice of the sets S and T!

Applications of Assouad's Lemma

Theorem (5)

For any $p, p', q \geq 1$ and $s > 0$, if L is a sufficiently large constant, then

$$\inf_{\tilde{\mu}} \sup_{f \in \mathcal{B}^s_{p',q}(L)} \mathbb{E} W_p(f, \tilde{\mu}) \gtrsim \begin{cases} n^{-\frac{1+s/p}{d+s}}, & d - s \geq 2p \\ n^{-\frac{1}{2p}}, & d - s < 2p \end{cases}$$

where the infimum is taken over all estimators $\tilde{\mu}$ based on n observations.

Proof.

1. if $d - s < 2p$:

Let $g_0 \in \mathcal{B}^s_{p',q}$ be supported in $[0, 1/3]^d$ and g_1 be a transition of g_0 supported on $[2/3, 1]^d$. For $\lambda \in [1, 1]$, define $f_\lambda := 2^{-1}[(1 + \lambda) g_0 + (1 - \lambda) g_1]$, then

$$\int (\sqrt{f_\lambda(x)} - \sqrt{f_{-\lambda}(x)})^2 dx = \int_{[0,1/3]^d} + \int_{[2/3,1]^d} \asymp (\sqrt{1+\lambda} - \sqrt{1-\lambda})^2 \asymp \lambda^2$$

Applications of Assouad's Lemma

(continued)
Let $S = [0, 1/3]^d$, $m = 1$ and $\{f_\lambda, f_{-\lambda}\} \subset \mathcal{B}^s_{p',q}(L)$, Lemma 5 implies that $W_p(f_\lambda, f_{-\lambda}) \gtrsim |f_\lambda(S) - f_{-\lambda}(S)|^{1/p} \asymp \lambda^{1/p}$, choosing λ such that $\lambda \asymp n^{-1/2}$ yields the claim.

2. if $d - s \geq 2p$:
According to the proposition derived from Assouad's Lemma above, we should construct as many distributions as possible to gain a large m that leads to a tight result under the condition that $H^2(\mu_\omega, \mu_{\omega'}) = O(n^{-1})$.

Let $g_0 \in \mathcal{B}^s_{p',q}$ supported in $[0, 1/3]^d$, $\Gamma = \{(\gamma_i^0, \gamma_i^1)\}_i$. For $\varepsilon \in \{0, 1\}^{|\Gamma|}$, define

$$f_\varepsilon := \sum_{1 \leq i \leq |\Gamma|} h(x - \gamma_i^{\varepsilon_i}) + \tau g_0, \quad \tau = 1 - |\Gamma|\delta > 0$$

here h has mini support and integral $\int h = \delta$, so as $h(x - \gamma_i^{\varepsilon_i})$, and all supports of $h(\cdot - \gamma_i^{\varepsilon_i}), g_0$ are well separated by at least c.

Applications of Assouad's Lemma

(continued)

The choice of ε differs f_ε, especially their supports. Let $\Delta(\varepsilon, \varepsilon') = \{i : \varepsilon_i \neq \varepsilon'_i\}$

$$S = \bigcup_{i \in \Delta(\varepsilon,\varepsilon')} \mathsf{supp}(h(x - \gamma_i^{\varepsilon_i})), \quad T = (\mathsf{supp}(f_\varepsilon) \cup \mathsf{supp}(f_{\varepsilon'})) \setminus S$$

hence $d(S,T) \geq c$, Lemma 5 and $\rho(\varepsilon, \varepsilon') \leq |\Gamma|$ imply

$$\begin{aligned} W_p(f_\varepsilon, f_{\varepsilon'}) &\geq c|\mu_\varepsilon(S) - \mu_{\varepsilon'}(S)|^{1/p} = c[\delta\rho(\varepsilon,\varepsilon')]^{1/p} \\ &\geq c[\delta \cdot |\Gamma|^{1-p}]^{1/p} \cdot \rho(\varepsilon, \varepsilon') = c\delta^{1/p}|\Gamma|^{1/p-1} \cdot \rho(\varepsilon,\varepsilon') \end{aligned}$$

When $\rho(\varepsilon, \varepsilon') = 1$, we have

$$\int (\sqrt{f_\varepsilon} - \sqrt{f_{\varepsilon'}})^2 \leq \int |f_\varepsilon - f_{\varepsilon'}| \lesssim \mu_\varepsilon(S) = \delta$$

Choosing $\delta \asymp n^{-1}$ then the proposition implies

$$\inf_{\tilde{\mu}} \sup_{f \in \mathcal{B}^s_{p',q}(L)} \mathbb{E} W_p(f, \tilde{\mu}) \gtrsim c\delta^{1/p}|\Gamma|^{1/p-1} \cdot |\Gamma| = c|\Gamma|^{1/p} \cdot n^{-1/p}$$

Applications of Assouad's Lemma

(continued)

We want to enrich the set Γ under the restriction that the supports are separated by at least c, to the upper bound: $|\Gamma| \lesssim (c+r)^{-d}$, where r is the diameter of supp(h), for example

$$\{\gamma_i^0\}_i \cup \{\gamma_i^1\}_i = \{(Kk_1(c+r), \cdots, Kk_d(c+r)) \notin [0,1/3]^d : (k_1, \cdots, k_d) \in \mathbb{Z}^d\}$$

the optimal choice should be $c \asymp r$: $c|\Gamma|^{1/p} \cdot n^{-1/p} \asymp c^{1-d/p} n^{-1/p}$.

We want to choose the smallest c to get the optimal results, while the conditions diam(supp(h)) $\asymp c$, $1 - |\Gamma|\delta > 0$ and $h \in \mathcal{B}^s_{p',q}(L)$ only allows $c \gtrsim n^{-1/(s+d)}$, which yields the final conclusion.

Actually the paper constructed $h = Mc^s g_0(x/c) \in \mathcal{B}^s_{p',q}(L)$ for some constant M, then $\delta = \int h = c^{s+d} \asymp 1/n$, diam(supp($h$)) $\asymp c \cdot$ diam(supp(g_0)) $\asymp c$, $|\Gamma| \asymp n^{\frac{d}{d+s}}$.

Applications of Assouad's Lemma

Theorem (?)

For $p \geq 1$ and $s > 0$, if $\sigma^2 > \max\{1, (d+s)/(2s)\}$ and L is a sufficiently large constant , then

$$\inf_{\tilde{\mu}} \sup_{f \in \mathcal{C}^s(L;\sigma^2)} \mathbb{E} W_p(f, \tilde{\mu}) \gtrsim (\log n)^{\frac{d}{2p}} n^{-\frac{1+s/p}{d+s}}$$

$\mathcal{C}^s(L;\sigma^2)$ stands for the set of probability densities on $\mathbb{R}^d$ with s-Hölder norm bounded by L that satisfy $\mathbb{E}_{X\sim f} \exp(\|X\|^2 / 2d\sigma^2) \leq 2$.

Proof.

We follow the same idea in Theorem 5, let $g_0 \in \mathcal{C}^s(L)$ supported in $[0, 1/3]^d$, $\Gamma = \{(\gamma_i^0, \gamma_i^1)\}_i$. For $\varepsilon \in \{0,1\}^{|\Gamma|}$, define

$$f_\varepsilon := \sum_{1 \leq i \leq |\Gamma|} h(x - \gamma_i^{\varepsilon_i}) + \tau g_0, \quad \tau = 1 - |\Gamma|\delta > 0$$

here $\int h \asymp n^{-1}$, and all supports of $h(\cdot - \gamma_i^{\varepsilon_i}), g_0$ are well separated by at least c.

Applications of Assouad's Lemma

(continued)
The difference is that, we limit h and $h(x-\gamma)$ to be supported on $[0,B]^d$, which also limits the size of $|\Gamma| \lesssim (B/c)^d$, then gradually increase B to ∞ as $n \to \infty$.

We still have $W_p(f_\varepsilon, f_{\varepsilon'}) \geq cn^{-1/p}|\Gamma|^{1/p-1} \cdot \rho(\varepsilon, \varepsilon')$ and $H^2(f_\varepsilon, f_{\varepsilon'}) \lesssim n^{-1}$ when $\rho(\varepsilon, \varepsilon') = 1$, hence

$$\inf_{\tilde{\mu}} \sup_{f \in \mathcal{C}^s(L;\sigma^2)} \mathbb{E} W_p(f, \tilde{\mu}) \gtrsim c|\Gamma|^{1/p} \cdot n^{-1/p} \gtrsim n^{-1/p} B^{d/p} c^{1-d/p}$$

In addition, here we need to ensure that $f_\varepsilon \in \mathcal{C}^s(L;\sigma^2)$:

$$\begin{aligned}\int e^{\frac{|x|^2}{2d\sigma^2}} f_\varepsilon(x)dx &\leq \int e^{\frac{|x|^2}{2d\sigma^2}} g_0(x)dx + \sum_{i \leq |\Gamma|} \int e^{\frac{|x|^2}{2d\sigma^2}} h(x-\gamma_i^{\varepsilon_i})dx \\ &\leq e^{1/18\sigma^2} + e^{B^2/2\sigma^2}|\Gamma| \int h \leq 2\end{aligned}$$

which gives a constraint on B and σ^2.

Applications of Assouad's Lemma

(continued)

Still letting $c \asymp n^{-\frac{1}{s+d}}$, $h = Mc^s g_0(x/c) \in \mathcal{C}^s(L)$. We want to ensure that

$$e^{B^2/2\sigma^2}|\Gamma| \int h \asymp e^{B^2/2\sigma^2} \cdot B^d n^{-\frac{s}{d+s}} \ll 1$$

This can be done by letting $\exp(B^2/2\sigma^2) \asymp n^\alpha$ with $\alpha < s/(d+s)$, i.e. $B^2 \asymp 2\alpha\sigma^2 \log n \asymp \log n$, then

$$\inf_{\tilde{\mu}} \sup_{f \in \mathcal{C}^s(L;\sigma^2)} \mathbb{E} W_p(f, \tilde{\mu}) \gtrsim n^{-1/p} B^{d/p} c^{1-d/p} \asymp (\log n)^{\frac{d}{2p}} n^{-\frac{1+s/p}{d+s}}$$

The condition $\sigma^2 > (d+s)/2s$ allowed B to be $\sqrt{\log n}$, it seemed that scaling B with appropriate coefficients that relies on given σ^2, d and s could remove this condition. Moreover, the condition $\sigma^2 > 1$ could also be removed, given the existence of compact supported density function in $\mathcal{C}^s(L;\sigma^2)$, i.e. the existence of g_0 (as L is sufficiently large).

Applications of Assouad's Lemma

(continued)
[1] [12][13] $\delta = \int h$

$$g_1 = (1-\delta)g_0 + h(x-\gamma), \qquad \gamma = \sqrt{\log n} \cdot \mathbf{1}$$

$$\left(1 - \frac{H^2(g_0, g_1)}{2}\right)^2 = \left(\int \sqrt{}\right)$$

$$H^2(g_0, g_1) = 2 - 2\int \sqrt{g_0[(1-\delta)g_0 + h]} = 2 - \sqrt{1-\delta}\int_S g_0 = 2[1 - \sqrt{1-\delta}]$$

$$H^2(g_0, g_1) = \int [\sqrt{(1-\lambda)g_0 + h} - \sqrt{(1-2\lambda)g_0 + 2h}]^2 dx = \int_S + \int_T$$

$$= (\sqrt{1-\lambda} - \sqrt{1-2\lambda})^2 + (\sqrt{2} - 1)\lambda$$

Table of Contents

References I

[1] J. Niles-Weed and Q. Berthet, "Minimax estimation of smooth densities in Wasserstein distance," *The Annals of Statistics*, vol. 50, no. 3, pp. 1519 – 1540, 2022. [Online]. Available: https://doi.org/10.1214/21-AOS2161

[2] C. Villani *et al.*, *Optimal transport: old and new*. Springer, 2009, vol. 338.

[3] N. Fournier and A. Guillin, "On the rate of convergence in wasserstein distance of the empirical measure," *Probability theory and related fields*, vol. 162, no. 3-4, pp. 707–738, 2015. [Online]. Available: https://doi.org/10.1007/s00440-014-0583-7

[4] N. Si, "Optimal transport in data science: Theory and applications." [Online]. Available: https://nian-si.github.io/ot.pdf

[5] J. Weed and F. Bach, "Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance," *Bernoulli*, vol. 25, no. 4A, pp. 2620 – 2648, 2019. [Online]. Available: https://doi.org/10.3150/18-BEJ1065

[6] A. Tsybakov, *Introduction to Nonparametric Estimation*, ser. Springer Series in Statistics. Springer New York, 2008.

References II

[7] W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov, *Wavelets, Approximation, and Statistical Applications*, ser. Lecture Notes in Statistics. Springer New York, 2012.

[8] H. P. Rosenthal, "On the span in lp of sequences of independent random variables. ii," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971)*, vol. 2, 1972, pp. 149–167.

[9] L. N. Lénaïc Chizat, "Université paris-saclay, lecture notes: Optimal transport," 2020. [Online]. Available: https://lchizat.github.io/ot2020orsay.html

[10] J.-D. Benamou and Y. Brenier, "A computational fluid mechanics solution to the monge-kantorovich mass transfer problem," *Numerische Mathematik*, vol. 84, pp. 375 – 393, 2000. [Online]. Available: https://doi.org/10.1007/s002110050002

References III

[11] S. Nietert, Z. Goldfeld, and K. Kato, "Smooth p-wasserstein distance: Structure, empirical approximation, and statistical applications," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 18–24 Jul 2021, pp. 8172–8183. [Online]. Available: https://proceedings.mlr.press/v139/nietert21a.html

[12] V. Divol, "Measure estimation on manifolds: an optimal transport approach," *Probability Theory and Related Fields*, vol. 183, pp. 581 – 647, 2022. [Online]. Available: https://doi.org/10.1007/s00440-022-01118-z

[13] B. Maury, A. Roudneff-Chupin, and F. Santambrogio, "A macroscopic crowd motion model of gradient flow type," *Mathematical Models and Methods in Applied Sciences*, vol. 20, 2010. [Online]. Available: https://doi.org/10.1142/S0218202510004799

[14] P. Assouad, "Deux remarques sur l'estimation," *Comptes Rendus des Séances de l'Académie des Sciences. Série I. Mathématique*, vol. 296, no. 23, pp. 1021–1024, 1983.

References IV

[15] C. Canonne, "Assouad and le cam proving distribution learning and testing lower bound," 2017. [Online]. Available: https://github.com/ccanonne/probabilitydistributiontoolbox/blob/master/assouadlemma.pdf

Definition (Push-forward and transport map)

$T : X \to Y$ is a measurable map between two metric spaces. The push-forward of $\mu \in \mathcal{P}(X)$ by T is the measure $\nu = T_{\#}\mu$ on Y defined by

$$\nu(B) = T_{\#}\mu(B) = \mu(T^{-1}(B))$$

for any Borel set B in Y. And such measurable map $T : X \to Y$ is called a transport map between μ and ν.

Definition (Monge Problem)

Consider two probability measures $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ on two metric spaces with a cost function $c : X \times Y \to \mathbb{R}_{\geq 0}$, minimize the cost

$$\mathrm{MP}(\mu, \nu) := \inf \left\{ \int_X c(x, T(x))\mu(dx) \mid T : X \to Y, T_{\#}\mu = \nu \right\}$$

Remark: such transport map $T : X \to X$ might not exist!

Definition (Transport plan)

A transport plan between two probabily measures $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$ is a probability measure $\pi \in \mathcal{P}(X \times Y)$, whose marginals are μ and ν. The space of transport plans is denoted $\Pi(\mu, \nu)$, i.e.

$$\Pi(\mu, \nu) = \{\pi \in \mathcal{P}(X \times Y) \mid \pi(A \times Y) = \mu(A),\ \pi(X \times B) = \nu(B)\}$$

If T be a transport map between μ and ν, and define $\pi_T = (\mathrm{id}, T)_{\#}\mu$. Then, π_T is a transport plan between μ and ν.

Definition (Kantorovich Problem)

Same conditions as Monge Probelm, but

$$\mathrm{KP}(\mu, \nu) := \inf \left\{ \int_X c(x, y)\pi(dxdy) \mid \pi \in \Pi(\mu, \nu) \right\}$$