Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces

JING LEI

Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
E-mail: jinglei@andrew.cmu.edu

We provide upper bounds of the expected Wasserstein distance between a probability measure and its empirical version, generalizing recent results for finite dimensional Euclidean spaces and bounded functional spaces. Such a generalization can cover Euclidean spaces with large dimensionality, with the optimal dependence on the dimensionality. Our method also covers the important case of Gaussian processes in separable Hilbert spaces, with rate-optimal upper bounds for functional data distributions whose coordinates decay geometrically or polynomially. Moreover, our bounds of the expected value can be combined with mean-concentration results to yield improved exponential tail probability bounds for the Wasserstein error of empirical measures under Bernstein-type or log Sobolev-type conditions.

Keywords: concentration inequality; empirical measure; empirical process; functional data; Wasserstein distance

1. Introduction

Let μ, ν be two probability measures on a Banach space $\mathcal{X}$, the *Wasserstein distance* of order p ($p \in [1, \infty)$) between μ, ν is

$$W_p(\mu, \nu) = \inf_{\xi \in \mathcal{C}(\mu, \nu)} \left[\mathbb{E}_{(X,Y) \sim \xi} \|X - Y\|^p \right]^{1/p},$$

where $\mathcal{C}(\mu, \nu)$ denotes the collection of all distributions on $\mathcal{X}^2$ with marginal distributions being μ and ν.

Wasserstein distances have a clear intuitive meaning: What is the minimum cost if we want to obtain ν by transporting the probability mass in μ? Here the cost is defined as the product of probability mass moved and the distance moved raised to the pth power. Therefore, the Wasserstein distance is also called "optimal transport distance" or "earth mover's distance". The problem of optimal transport can be traced back to [30] and [24].

Since the introduction in [38], the Wasserstein distances have become an important tool in computer vision and statistical machine learning. In addition to the connection with optimal transport, Wasserstein distances have some desirable features. For example, they can be meaningfully defined for any two distributions without any requirement on the existence of density or

absolute continuity. Other measurements, such as the Kullback-Leibler divergence, have more stringent requirements on μ and ν. See [33] for a more thorough historical review of Wasserstein distances and their applications in statistics, and [40] for further details about Wasserstein distances and optimal transport in a broader context.

In statistics and machine learning, we often do not have access to μ but only its empirical version $\hat{\mu}$, which puts $1/n$ mass at each one of n independent samples from μ. A fundamental problem is to understand $W_p(\hat{\mu}, \mu)$. When $p = 1$ the Kantorovich-Rubinstein duality [23] implies that $W_1(\hat{\mu}, \mu)$ is equivalent to the supremum of the empirical process indexed by Lipschitz-1 functions. As a consequence, [13] provides sharp lower and upper bounds of $\mathbb{E}W_1(\hat{\mu}, \mu)$ for μ supported on a bounded finite dimensional set. [34] studies the case when μ is the uniform distribution on a d-dimensional unit cube. For general distributions, [7,12,16] establish sharp upper bounds of $\mathbb{E}W_p(\hat{\mu}, \mu)$ in finite dimensional Euclidean spaces. Recently, [32,41] provide similar results for general distributions on a bounded metric space $\mathcal{X}$.

In this paper, we study upper bounds of $\mathbb{E}W_p(\hat{\mu}, \mu)$ for distributions in unbounded Banach spaces, including higher dimensional Euclidean spaces ($\mathcal{X} = \mathbb{R}^d$ where d may grow to ∞ as n increases) and separable Hilbert spaces (square-integrable functions on $[0, 1]$) as important examples. Our argument combines the strengths from several recent works and provides more general and improved results in settings of practical interest. In addition to bounding $\mathbb{E}W_p(\hat{\mu}, \mu)$, we also establish a concentration inequality for $W_p(\hat{\mu}, \mu)$ around its expected value, which is dimension-free and offers improvements over existing results.

1.1. Motivating examples

The convergence of empirical measure in Wasserstein distances in high-dimensional and functional spaces can be useful in different machine learning and statistical inference contexts. Here we describe two examples in detail.

Example 1 (Unsupervised deep learning). The first example is the Wasserstein GAN [3], where GAN stands for "generative adversarial network" [18], a flexible framework for unsupervised learning. The goal of GAN is to approximate an unknown distribution μ_r on a sample space $\mathcal{X}$, given an i.i.d. sample from μ_r. A typical example is the hypothetical distribution of all images. To this end, GAN starts from a simple underlying distribution μ_z on a possibly different space $\mathcal{Z}$, such as the standard multivariate Gaussian in $\mathcal{Z} = \mathbb{R}^{100}$, and then finds a mapping $g : \mathcal{Z} \mapsto \mathcal{X}$, such that the distribution induced by $g(Z)$ on $\mathcal{X}$, denoted by μ_g, is close to μ_r.

In practice, neither μ_r nor μ_g is available. But the empirical version $\hat{\mu}_r$ is given as the data, and $\hat{\mu}_g$ can be generated by passing i.i.d. samples from μ_z through the black-box function g. So one has to work with the empirical distributions instead of the original underlying distributions. In particular, Wasserstein GAN tries to find a mapping g to minimize $W_1(\hat{\mu}_g, \hat{\mu}_r)$ using deep neural networks. We refer to the original papers [3,18] for the detailed realization and implementation of such a minimization problem. Since its invention, Wasserstein GANs have found great success in image data due to the nice behavior of Wasserstein distances between mutually singular distributions.

However, in machine learning, we are ultimately interested in the ability to approximate/predict future observations generated from the same underlying distribution. This leads

to the problem of *generalization*. In the context of Wasserstein GAN, this problem further reduces to controlling $W_1(\mu_g, \mu_r)$. Given that the WGAN algorithm finds a g such that $W_1(\hat{\mu}_g, \hat{\mu}_r)$ is small, one would know $W_1(\mu_g, \mu_r)$ is small if both $W_1(\mu_g, \hat{\mu}_g)$ and $W_1(\mu_r, \hat{\mu}_r)$ are small. Since WGAN is typically applied to images and videos, it is desirable to cover cases where $\mathcal{X}$ is high-dimensional or infinite-functional.

Example 2 (Latent space representation of exchangeable random graphs). Exchangeable random graphs [2,21,22] have been an important probabilistic tool in network research [4,28]. The existing parameterization of such exchangeable random graphs, called the *graphon* [28], has some undesirable identifiability issues, making the corresponding statistical inferences challenging [1,17,43]. Recently, [27] established a new parameterization for exchangeable random graphs to facilitate the statistical inferences for network data. There are three findings in [27].

 (i) Under certain mild regularity conditions, an exchangeable random graph $\mathcal{A}$ can be parameterized by a distribution $\mu_{\mathcal{A}}$ on a separable Hilbert space $\mathcal{H}$.
 (ii) For two exchangeable random graphs $\mathcal{A}_1$ and $\mathcal{A}_2$ with corresponding Hilbert space distribution parameters μ_1 and μ_2, the difference in graph limits between $\mathcal{A}_1$ and $\mathcal{A}_2$ (also known as the cut-distance) can be upper bounded by $W_1(\mu_1, \mu_2)$.
(iii) Suppose we observe a partial realization of $\mathcal{A}$ with n nodes, then each node can be embedded in the Hilbert space $\mathcal{H}$ as an i.i.d. sample from $\mu_{\mathcal{A}}$, and it is possible to find a good approximation of $\hat{\mu}_{\mathcal{A}}$ in Wasserstein-1 distance using the observed random graph with n nodes. Here $\hat{\mu}_{\mathcal{A}}$ is the empirical version of $\mu_{\mathcal{A}}$ with $1/n$ probability mass at each of the embedded nodes.

Combining the above three claims, one would be able to recover $\mu_{\mathcal{A}}$ in Wasserstein-1 distance from a partial realization of the exchangeable random graph $\mathcal{A}$ if one can show that $W_1(\hat{\mu}_{\mathcal{A}}, \mu_{\mathcal{A}})$ is small. Again, given the context we wish to consider this problem in an infinite-dimensional separable Hilbert space without boundedness assumptions.

1.2. Overview of results

In Section 2, we develop a general argument for bounding $\mathbb{E}W_p(\hat{\mu}, \mu)$. This approach combines the hierarchical transport program under moment conditions developed in [12,16], with the abstract partition approach in [41]. Comparing to [12,16], both of which focus on fixed dimensionality, our argument uses covering numbers of a baseline set that is compatible with the corresponding moment condition, leading to improved rates of convergence in high dimensional settings, as well as applicability to general Banach spaces. Comparing to [32,41], both of whom focus on bounded spaces, our result extends to the unbounded case using a moment condition. The result of [41] is for fixed dimensionality, while the infinite dimensional example in [32] is a special case of the moment condition used in this paper. An example that cannot be covered by these existing works is infinite dimensional Gaussian processes.

In Section 3, we demonstrate the advantage of abstract partition in Euclidean spaces with large dimensionality. If $\mathcal{X} = \mathbb{R}^d$ and $\mathbb{E}_{X \sim \mu} \|X\|^q \leq M_q^q$ for some $p < q \ll d \ll \log n$. Then (see

Theorem 3.1)

$$\mathbb{E}W_p(\hat{\mu}, \mu) \le c_{p,q} M_q n^{-1/d},$$

where the constant $c_{p,q}$ depends only on (p, q) but not d. In contrast, the results in [12,16,32,41] have a constant as an unbounded function of d.

In Section 4, we further apply the general recipe to distributions on separable Hilbert spaces under ellipsoid-type moment conditions. Let μ be a distribution on $\mathcal{X} = \{x \in \mathbb{R}^\infty : \sum_{m \ge 1} x_m^2 < \infty\}$. Let $\tau = (\tau_m : m \ge 1)$ be a sequence of nonnegative numbers and $\rho_\tau(x) = [\sum_{m \ge 1} (x_m/\tau_m)^2]^{1/2}$. The moment condition we consider here is $\mathbb{E}\rho_\tau^q(X) = M_q^q < \infty$. We consider two popular cases of τ in functional data analysis. In the first case, $\tau_m = m^{-b}$ for some $b > 1/2$, which corresponds to a polynomial decay of the coordinates. In this case, we have (see Theorem 4.1), for some constant $c_{p,q,b}$ depending only on (p, q, b),

$$\mathbb{E}W_p(\hat{\mu}, \mu) \le c_{p,q,b} M_q (\log n)^{-b}.$$

In the second case $\tau_m = \gamma^{-m}$ for some $\gamma > 1$, which corresponds to an exponential decay of the coordinates. In this case we have (Theorem 4.2), for some constant $c_{p,q,\gamma}$ depending only on (p, q, γ),

$$\mathbb{E}W_p(\hat{\mu}, \mu) \le c_{p,q,\gamma} M_q e^{-\sqrt{2 \log \gamma \log n}}.$$

These upper bounds are rate-optimal because in each case we can construct a corresponding μ such that the upper bound is matched by a lower bound with the same rate.

In Section 5, we establish concentration inequalities for $W_p(\hat{\mu}, \mu)$. Assuming a sub-exponential tail behavior of $\|X\|$, the rate of concentration of $W_p(\hat{\mu}, \mu)$ around its expected value can be as small as $n^{-1/2}$ when $p = 1$, which can vanish faster than the optimal upper bound of $\mathbb{E}W_p(\hat{\mu}, \mu)$ in high-dimension settings. The new concentration inequality is dimension-free. In the case of $p = 1$, it provides a better concentration rate compared to [6] and [16] under the sub-exponential tail condition. We also extend an argument of [5,19] from the case of $\mathcal{X} = \mathbb{R}^1$ to more general cases. Our concentration results are based on a variation of McDiarmid's inequality and the Lipschitz property of Wasserstein distances.

Notation

$\mathbb{B}_x^r = \{y \in \mathcal{X} : \|x - y\| \le r\}$ is the ball of radius r centered at x.

$\mathbb{N}$ is the set of all positive integers.

For a random variable Z, $\|Z\|_p = (\mathbb{E}|Z|^p)^{1/p}$ denotes the L_p norm of Z.

$\|x\|$ denotes the norm of x in the Banach space $\mathcal{X}$.

L_2 denotes the separable Hilbert space $\{x \in \mathbb{R}^\infty : \sum_{m \ge 1} x_m^2 < \infty\}$.

For $B \subseteq \mathcal{X}$ and $a \in \mathbb{R}$: $aB = \{ax : x \in B\}$.

For a set B and a measure μ, $\mu|_B$ is the confinement of μ on B: $(\mu|_B)(A) = \mu(B \cap A)$.

2. A general upper bound

Given a distribution μ on an unbounded space $\mathcal{X}$ and its empirical version $\hat{\mu}$ of sample size n. The Wasserstein distance $W_p(\hat{\mu}, \mu)$ is determined by essentially two aspects of the problem: the

complexity of the support of μ and the tail behavior of μ. We describe these two aspects in detail and introduce the corresponding techniques.

2.1. Covering and partition in $\mathcal{X}$

The first quantity that affects $W_p(\hat{\mu}, \mu)$ comes from the dimensionality (or complexity) of the space $\mathcal{X}$. Suppose for now that μ is supported on a bounded subset $B_0 \subset \mathcal{X}$. The complexity of B_0 can be described by its ϵ-covering number.

Definition 1 (ϵ-covering). For $B \subseteq \mathcal{X}$, and $\epsilon \in (0, 1]$, the ϵ-covering number, denoted as $N_\epsilon(B)$ is

$$N_\epsilon(B) = \inf\left\{ N \in \mathbb{N} : \exists x_1, \ldots, x_N \in B, \text{ s.t. } B \subseteq \bigcup_{i=1}^{N} \mathbb{B}_{x_i}^\epsilon \right\}.$$

$\log N_\epsilon(B)$ is called the ϵ-entropy number of B and the speed at which it increases as ϵ decreases plays an important role in the empirical process theory.

Given an $\epsilon \in (0, 1]$, let $x_1, \ldots, x_{N_\epsilon(B_0)}$ be the centers of an ϵ-covering of B_0. Let $F_1 = \mathbb{B}_{x_1}^\epsilon \cap B_0$, and $F_i = (\mathbb{B}_{x_i}^\epsilon \cap B_0)\backslash(\bigcup_{1 \leq j < i} F_j)$ for $2 \leq i \leq N_\epsilon(B_0)$, then $\mathcal{A} = \{F_i : 1 \leq i \leq N_\epsilon(B_0)\}$ is a partition of B_0 satisfying $\max_{F \in \mathcal{A}} \text{diam}(F) \leq 2\epsilon$ and $\text{card}(A) \leq N_\epsilon(B_0)$.

A basic tool for controlling $W_p(\hat{\mu}, \mu)$ is to construct a sequence of nested ϵ_ℓ-partitions for a sequence of geometrically decaying $(\epsilon_\ell : \ell \in 0 \cup \mathbb{N})$. [12,16] use $B_0 = [-1, 1]^d$ and recursive dyadic partitions of B_0. The use of general ϵ-partitions first appeared in [41].

In our proof of upper bound, we consider $\epsilon_\ell = 3^{-(\ell+1)}$ as in [41] and define

$$\bar{N}_\ell(B_0) = N_{3^{-(\ell+1)}}(B_0), \quad \ell \in 0 \cup \mathbb{N}. \tag{1}$$

The complexity of B_0 will be reflected in the upper bound of $W_p(\hat{\mu}, \mu)$ through $\bar{N}_\ell$.

The following lemma is a basic result in upper bounding $W_p(\mu, \nu)$ in the bounded support case using a hierarchical partition.

Lemma 2.1. *Let μ and ν be two probability measures supported on B_0. Let $\ell^* \in \{0\} \cup \mathbb{N}$ and $\{B_0\} = \mathcal{A}_0, \mathcal{A}_1, \ldots, \mathcal{A}_{\ell^*}$ be a sequence of nested partitions of B_0 such that $\sup_{F \in \mathcal{A}_\ell} \text{diam}(F) \leq c3^{-\ell}$ for some constant c, then there exists a constant c_p depending only on p such that*

$$W_p^p(\mu, \nu) \leq c_p\left(3^{-p\ell^*} + \sum_{\ell=0}^{\ell^*} 3^{-p\ell} \sum_{F \in \mathcal{A}_\ell} |\nu(F) - \mu(F)| \right).$$

Lemma 2.1 first appears in its general form as Proposition 1 of [41], which extends the Euclidean space version with dyadic hypercube partition in [12,16]. In Appendix B, we give an alternative proof of it, adapting the elegant Markov chain argument in [12].

In Lemma 2.1, the sum over ℓ comes from the fact that the best way to transport the mass in μ to $\hat{\mu}$ is through its blurred versions on the ϵ_ℓ-partition. Let $\mathcal{A}_\ell$ be the partition of B_0 at level ℓ, the blurred version, μ_ℓ, of μ is (using the convention $0/0 = 0$)

$$\mu_\ell = \sum_{F \in \mathcal{A}_\ell} \mu_\ell|_F, \quad \text{with } \mu_\ell|_F = \frac{\hat{\mu}(F)}{\mu(F)} \mu|_F.$$

These blurred versions can be used to construct a chain of transports that link between μ and $\hat{\mu}$, while each having small transport cost.

2.2. Telescope decomposition and moment condition

Now we drop the assumption that μ is supported on a bounded set B_0. The second quantity that affects $W_p(\hat{\mu}, \mu)$ comes from the tail behavior of μ, which we control using a "telescope decomposition" of $\mathcal{X}$, combined with a moment condition.

Let $\rho : \mathcal{X} \mapsto \{0\} \cup \mathbb{R}^+$ be a homogenous functional satisfying

(R1) $\rho(ax) = |a|\rho(x)$ for all $a \in \mathbb{R}$;
(R2) $B_{0,\rho} \stackrel{\text{def}}{=} \{x : \rho(x) \leq 1\} \subseteq \mathbb{B}_0^1$.

Condition (R1) simply requires homogeneity of ρ. Condition (R2) ensures that ρ dominates the Banach space norm.

When $\mathcal{X} = \mathbb{R}^d$, an example of ρ is $\rho(x) = \|x\|$ and the corresponding $B_{0,\rho}$ is the unit Euclidean ball. When $\mathcal{X} = L_2$, an example of ρ is $\rho_\tau(x) = [\sum_{m \geq 1}(x_m/\tau_m)^2]^{1/2}$ for some $\tau \in (\mathbb{R}^+)^\infty$ satisfying $\tau_m \leq 1$ for all m.

Sometimes the dimensionality of $\mathbb{B}_0^1$ is too high to yield any meaningful upper bound. For example, if $\mathcal{X} = L_2$ then $N_\epsilon(\mathbb{B}_0^1) = \infty$ for all values of ϵ smaller than a constant. However, $B_{0,\rho}$ can be substantially smaller than $\mathbb{B}_0^1$ for appropriate choices of ρ, so that we can still obtain useful upper bounds.

To deal with unbounded support of μ, we decompose μ to its confinements on a sequence of "telescoping" subsets $(B_j : j \in 0 \cup \mathbb{N})$, where $B_0 = B_{0,\rho}$ and

$$B_j = \left(2^j B_0\right)\backslash B_{j-1}, \quad \forall j \in \mathbb{N}.$$

The tail behavior of μ is characterized by $\mu|_{B_j}$ for large values of j. Our analysis will consider $\tilde{\mu}_j$, the induced measure of $\frac{\mu|_{B_j}}{\mu(B_j)}$ after the mapping $x \mapsto 2^{-j}x$, and apply the bounded support results to $\tilde{\mu}_j$.

The following "telescoping lemma" generalizes the corresponding argument in [16] to the case of general Banach spaces and general moment conditions.

Lemma 2.2. *Let ρ be a function that satisfies* (R1) *and* (R2). *Let μ and ν be two probability measures on $\mathcal{X}$ such that $\rho(x) < \infty$ almost surely under μ and ν. For $j \geq 0$, define $\tilde{\mu}_j$ ($\tilde{\nu}_j$) to be*

the normalized probability measure induced by $\mu|_{B_j}$ ($\nu|_{B_j}$) after the mapping $x \mapsto 2^{-j}x$. Then there exists a constant c_p depending only on p such that

$$W_p^p(\mu, \nu) \leq c_p \sum_{j \geq 0} 2^{pj}\left[\left(\mu(B_j) \wedge \nu(B_j)\right)W_p^p(\tilde{\mu}_j, \tilde{\nu}_j) + \left|\mu(B_j) - \nu(B_j)\right|\right]. \qquad (2)$$

The exponentially increasing factor 2^{pj} in the above sum needs to be offset by the small terms $\mu(B_j) \wedge \nu(B_j)$. This motivates the following moment condition:

$$\left\|\rho(X)\right\|_q = M_q < \infty \qquad (3)$$

for some constant $q > p$.

If the moment condition (3) holds with $M_q = 1$ for $X \sim \mu$, then the contribution of $W_p^p(\tilde{\mu}_j, \tilde{\nu}_j)$ in (2) becomes less important as j gets larger, because $\mu(B_j) \leq 2^{-q(j-1)}$.

The moment condition $\|\rho(X)\|_q = M_q < \infty$ for $X \sim \mu$ serves another purpose: It implies that $\rho(x) < \infty$ μ-almost surely. In this paper, we are mainly interested in the case $\nu = \hat{\mu}$, whose support is no larger than that of μ. As a result, we have $\rho(x) < \infty$ almost surely under μ and ν, as required by Lemma 2.2. An important consequence is that $\{B_j : j \geq 0\}$ forms a partition of the supports of μ and ν.

2.3. A general upper bound

Theorem 2.3. *Let $\bar{N}_\ell$ be defined as in (1) with $B_0 = B_{0,\rho}$ for a function ρ satisfying conditions R1 and R2. Assume the moment condition (3) holds with $q > p \geq 1$ for $X \sim \mu$, then for any $\ell^* \in \mathbb{N}$, we have*

$$\mathbb{E}W_p^p(\hat{\mu}, \mu) \leq c_{p,q} M_q^p \sum_{j \geq 0} 2^{pj}\left\{2^{-qj}3^{-p\ell^*} + \sum_{\ell=0}^{\ell^*} 3^{-p\ell}\left[2^{-qj} \wedge \left(\bar{N}_\ell 2^{-qj}n^{-1}\right)^{1/2}\right]\right\}. \qquad (4)$$

Theorem 2.3 looks similar to its counterparts in [12,16,32,41], and has combined advantages from them. Comparing to [12,16], (4) uses a finite sum over ℓ which allows for the use of customized ϵ-partitions. We shall see in Section 3 that this can improve the dependence on the Euclidean dimensionality d when $\mathcal{X} = \mathbb{R}^d$. This also allows us to cover the case of $\mathcal{X} = L_2$ with appropriately chosen ρ functions, as detailed in Section 4. On the other hand, [32,41] do not have the outer sum over j and only have the term for $j = 0$, which only works for bounded support. We prove Theorem 2.3 in Appendix B.

The details in (4) reflect the trade-off between the depth of the partition (indexed by ℓ), the sample size n, and the importance of each telescoping layer B_j. The most intriguing part of (4) is the term $2^{-qj} \wedge (\bar{N}_\ell 2^{-qj}n^{-1})^{1/2}$, where the "$\wedge$" comes from the fact $\mathbb{E}|\mu(F) - \hat{\mu}(F)| \leq (2\mu(F)) \wedge (\mu(F)/n)^{1/2}$ for any F. It suggests that for different layers B_j, the critical partition levels ℓ_j^* (the value of ℓ such that $2^{-qj} \asymp (\bar{N}_\ell 2^{-qj}n^{-1})^{1/2}$) will be different.

3. Euclidean spaces: Optimal dependence on d

In this section, we assume $\mathcal{X} = \mathbb{R}^d$, together with $\rho(x) = \|x\|$, the Euclidean norm.

Theorem 3.1. *If $\mathcal{X} = \mathbb{R}^d$ and $\mathbb{E}(\|X\|^q) = M_q^q < \infty$ for $X \sim \mu$ and some $q > p \geq 1$, then*

$$\mathbb{E}W_p(\hat{\mu}, \mu) \leq c_{p,q} M_q n^{-[\frac{1}{(2p)\vee d} \wedge (\frac{1}{p} - \frac{1}{q})]} (\log n)^{\zeta_{p,q,d}/p}, \tag{5}$$

where $c_{p,q}$ is a constant depending only on (p,q) (not d), and

$$\zeta_{p,q,d} = \begin{cases} 2 & \text{if } d = q = 2p, \\ 1 & \text{if "}d \neq 2p \text{ and } q = \dfrac{dp}{d-p} \wedge 2p\text{" or "}q > d = 2p\text{",} \\ 0 & \text{otherwise.} \end{cases}$$

Remark 1. Theorem 3.1 can be compared with Theorem 1 of [16] and Example 2 of [32] (the latter assumes bounded support, which corresponds to the case of $q = \infty$). The key difference is that our constant $c_{p,q} M_q$ does not depend on d, while in all existing similar results the constants depend on d in an unbounded manner. Such an improved dependence on d makes a qualitative difference when we want to project the infinite dimensional sample points onto a subspace whose dimension increases as the sample size n. This is particularly the case in the latent space network representation example described in Section 1.1.

Remark 2. The logarithm factor in (5) is active only in the boundary cases such as $d = 2p$ or $q = 2p$, and is upper bounded by $(\log n)^{2/p}$ in the worst case. For the cases without such a logarithm factor, the optimality of the main factor $n^{-[\frac{1}{(2p)\vee d} \wedge (\frac{1}{p} - \frac{1}{q})]}$ has been discussed in detail in [16].

We give the proof of Theorem 3.1 in Appendix C. To highlight the novel ingredients in our proof, here we consider the case $d > 2p$.

Suboptimality of hypercube partition. If we follow the proof of [16] that uses $B_0 = [-1, 1]^d$ and $\mathcal{A}_\ell$ as dyadic partition of B_0 with hypercubes of side length $2^{1-\ell}$, then the constant factor in (5) will have a factor of $\sqrt{d}$ that comes from the diameter of B_0.

Covering numbers of the unit Euclidean ball. To avoid such a $\sqrt{d}$ factor, we can choose $B_0 = \mathbb{B}_0^1$, the d-dimensional unit ball centered at 0. However, the best upper bound for the covering number of $\mathbb{B}_0^1$ is $N_\epsilon(\mathbb{B}_0^1) \leq \epsilon^{-d}(d+1)^c$ for some constant $c \geq 3/2$ [39], Theorem 3.1. Thus, we need to work with the upper bound $\bar{N}_\ell \leq 3^{(\ell+c_0)d}$ for some different positive constant $c_0 > 1$. Directly factoring out the $3^{c_0 d/2}$ factor from the $\bar{N}_\ell^{1/2}$ term in (4) will bring the undesirable (and indeed unnecessary) $3^{c_0 d/2}$ factor to the final upper bound.

Our proof avoids this by directly keeping track of each $\bar{N}_\ell$. The effect of d shows up differently in (4) for small and large value of j's, which are controlled by different techniques. Here

we briefly describe the case when j is small. When $d > 2p$, for small values of j the main contributing term in the inner sum of (4) is $3^{-p\ell}\bar{N}_\ell^{1/2}2^{-qj/2}n^{-1/2}$, which is an increasing geometric sequence and hence bounded by the last term. The key observation here is that the last term in this sequence is bounded by $3^{-p\ell}2^{-qj}$ by construction, and hence there is no need to use the detailed formula for $\bar{N}_\ell$ as a function of (ℓ, d). This is indeed a benefit of the abstract partition scheme where the diameter of covering balls is dimension-free.

4. Functional spaces

In this section, we consider the separable Hilbert space $\mathcal{X} = L_2 = \{x \in \mathbb{R}^\infty : \sum_{m=1}^\infty x_m^2 < \infty\}$. This functional space is isomorphic to $L_2([0, 1])$, the space of square integrable functions on $[0, 1]$, which is a commonly assumed data space in functional data analysis.

Now let μ be a probability measure on L_2. As discussed in Section 2.2, we can use an ellipsoidal moment condition on μ to deal with the infinite dimensionality. The function ρ used for our moment conditions has the following form

$$\rho_\tau(x) = \left[\sum_{m=1}^\infty \left(\frac{x_m}{\tau_m}\right)^2\right]^{1/2},$$

where $\tau = (\tau_m : m \geq 1)$ is a sequence of positive numbers with $\sup_m \tau_m \leq 1$.

The moment condition (3) becomes stronger if τ_m's are small. We consider two types of such τ sequences, which correspond to different speeds at which τ_m vanishes. Relevance of these scenarios in functional data analysis is discussed in Section 4.3.

We adopt a minimax perspective to illustrate the optimality of the upper bounds. Consider a class of distributions with a specified tail behavior, if we can show that a lower bound of the same rate is achieved by at least one member in this class, then the upper bound rate cannot be improved uniformly within such a class of distributions.

4.1. Polynomial decay

In this case, we consider τ sequences with

$$\tau_m = m^{-b}, \quad \forall m \geq 1 \tag{6}$$

for some $b > 1/2$.

The class of distributions of interest are those satisfy the moment condition (3) with $\rho = \rho_\tau$ and τ given in (6). In this case $\rho_\tau(x) = [\sum_m (m^b x_m)^2]^{1/2}$. Formally, define the distribution class

$$\mathcal{P}_{\text{poly}}(q, b, M_q) := \left\{\mu : \mathbb{E}_{X\sim\mu}\left[\sum_{m=1}^\infty (m^b X_m)^2\right]^{\frac{q}{2}} \leq M_q^q\right\}. \tag{7}$$

Theorem 4.1. *If p, q, b are constants such that $1 \le p < q$ and $b > 1/2$, then there exist positive constants $\underline{c}_{p,q,b}, \bar{c}_{p,q,b}$ depending on (p, q, b) such that*

$$\underline{c}_{p,q,b} M_q (\log n)^{-b} \le \sup_{\mu \in \mathcal{P}_{\mathrm{poly}}(q,b,M_q)} \mathbb{E} W_p(\hat{\mu}, \mu) \le \bar{c}_{p,q,b} M_q (\log n)^{-b}.$$

The poly-log rate in Theorem 4.1 comes from the fact that the metric entropy (logarithm of the covering number) of B_{0,ρ_τ} has a polynomial dependence on ϵ^{-1} and hence an exponential dependence on ℓ if $\epsilon = 3^{-(\ell+1)}$. This is in sharp contrast with the Euclidean case, where the metric entropy is a linear function of $\log(\epsilon^{-1})$ and hence a linear function of ℓ if $\epsilon = 3^{-(\ell+1)}$.

One can compare Theorem 4.1 with Example 4 of [32], which independently provides similar matching upper and lower bounds for a different function class. To make them directly comparable, we consider a typical special case where $\mathcal{X}$ consists of Lipschitz functions on $[0, 1]$ satisfying $f(0) = f(1)$. Such an $\mathcal{X}$ is a Hölder class, so example 4 of [32] suggests an upper bound of order $(\log n)^{-1}$. On the other hand, the Hölder class is a subset of $\mathcal{P}_{\mathrm{poly}}(q, b, M_q)$ with $b = 1$ under the trigonometric basis, which is a Sobolev ellipsoid (see Chapter 1.7.1 of [35]). Therefore, Theorem 4.1 also implies an upper bound of order $(\log n)^{-1}$.

4.2. Exponential decay

In this case, we consider τ sequences with exponential decay:

$$\tau_m = \gamma^{-(m-1)}, \quad \forall m \ge 1 \tag{8}$$

for some $\gamma > 1$.

The corresponding class of distributions satisfy the moment condition (3) with $\rho = \rho_\tau$ and τ given in (8). In this case, $\rho_\tau(x) = [\sum_m (\gamma^{m-1} x_m)^2]^{1/2}$. Define the distribution class

$$\mathcal{P}_{\mathrm{exp}}(q, \gamma, M_q) := \left\{ \mu : \mathbb{E}_{X \sim \mu} \left[\sum_{m=1}^{\infty} (\gamma^{m-1} X_m)^2 \right]^{\frac{q}{2}} \le M_q^q \right\}. \tag{9}$$

Theorem 4.2. *If p, q, γ are constants such that $1 \le p < q$ and $\gamma > 1$, then there exist positive constants $\underline{c}_{p,q,\gamma}, \bar{c}_{p,q,\gamma}$ depending only on (p, q, γ) such that*

$$\underline{c}_{p,q,\gamma} M_q e^{-\sqrt{2 \log \gamma \log n}} \le \sup_{\mu \in \mathcal{P}_{\mathrm{exp}}(q,\gamma,M_q)} \mathbb{E} W_p(\hat{\mu}, \mu) \le \bar{c}_{p,q,\gamma} M_q e^{-\sqrt{2 \log \gamma \log n}}.$$

The lower and upper bounds in Theorem 4.2 seem to be new in the literature. Theorem 4.2 shows that, even when the coordinates decay exponentially, the functional space is still fundamentally different from Euclidean spaces. As we will see in Lemma 4.3 below, the metric entropy is a quadratic, instead of linear, function of $\log(\epsilon^{-1})$, so we can only have a sub-polynomial rate of convergence in Theorem 4.2. Moreover, the exponential decay makes the problem qualitatively different from the polynomial decay, as this sub-polynomial rate is faster than the poly-log rate in Theorem 4.1.

In both Theorems 4.1 and 4.2, our proofs reflect that when $\mathcal{X}$ has higher dimensionality than Euclidean space, $W_p(\hat{\mu}, \mu)$ is determined by the value ϵ such that $N_\epsilon(B_0) \asymp n$.

The lower and upper bounds of the covering numbers $N_\epsilon(B_0)$ and $\bar{N}_\ell$ required in the proof of Theorem 4.2 are provided in the following lemma, which adapts the covering number bounds for finite dimensional ellipsoids in [14,15] to infinite dimensional ellipsoids using a truncation argument. Equivalent forms of this result have appeared in the literature (see, for example, [42], Theorem 7). The version presented here is more direct applicable for our purpose.

Lemma 4.3. *Let $B_0 = \{x \in L_2 : \rho_\tau(x) \leq 1\}$ with τ given in (8). Then for $\epsilon \in (0, 1]$,*

$$\log N_\epsilon(B_0) \geq \frac{[\log(\epsilon^{-1})]^2}{2 \log \gamma}$$

and for $\ell \in \mathbb{N}$,

$$\log \bar{N}_\ell = \log N_{3^{-(\ell+1)}}(B_0) \leq c_\gamma (\ell + c_1)^2 + \log c_0$$

with $c_\gamma = \frac{(\log 3)^2}{2 \log \gamma}$ and c_0, c_1 positive constants depending on γ.

4.3. Functional principal components

In the previous two subsections, we see that $W_p(\hat{\mu}, \mu)$ tends to be smaller if the majority probability mass of μ can be covered by an ellipsoid $B_0 = B_{0, \rho_\tau}$, with fast decaying axes τ_m. Intuitively speaking, this corresponds to an orthogonal basis of L_2, denoted as $\{\phi_m : m \geq 1\}$ such that the variability of the coordinate projections $\langle X, \phi_m \rangle$ decays fast when $X \sim \mu$. Such an explanation suggests an immediate connection to functional principal components analysis [20,31], where one can choose $\{\phi_m : m \geq 1\}$ to be the eigenvectors of the covariance operator of μ.

To facilitate discussion, we assume that $\mathbb{E}_{X \sim \mu} X = 0$ and $\mathbb{E}_{X \sim \mu} \|X\|^2 < \infty$. Then we can write the random function (in a separable Hilbert space, such as L_2) $X \sim \mu$ in its Karhunen–Loève decomposition

$$X = \sum_{m=1}^{\infty} \sigma_m Z_m \phi_m, \tag{10}$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ are the standard deviations of the coordinate projections of X ranked in decreasing order, and $(Z_m : m \geq 1)$ are uncorrelated random variables with mean 0 and unit variance. Standard theory shows that such a basis $(\psi_m : m \geq 1)$ leads to the fastest decay of the tail squared sum $\sum_{j \geq m} \sigma_j^2$ for all m.

The speed at which σ_m decays as m increases measures the "regularity" of the functional data distribution μ. Here we can make explicit connections between the eigen-decay of μ and the ellipsoidal moment conditions considered in the previous two subsections.

Proposition 4.4. *If the standardized principal component scores $(Z_m : m \geq 1)$ of μ satisfy $\sup_m \|Z_m\|_q = M < \infty$ with $q \geq 2$, then the following hold.*

1. (*Polynomial decay*) If

$$\sigma_m \leq c_0 m^{-(b_0 + 1/2)}$$

for some constants $b_0 > 1/2$, $c_0 > 0$, then the moment condition (3) holds with τ sequence given in (6) for any constant $b \in (1/2, b_0)$ and $M_q = c_{b, b_0, q} c_0 M$, where $c_{b, b_0, q}$ is a constant depending only on (b, b_0, q).

2. (*Exponential decay*) If

$$\sigma_m \leq c_0 \gamma_0^{-(m-1)}$$

for some constants $\gamma_0 > 1$, $c_0 > 0$, then the moment condition (3) holds with τ sequence given in (8) for any $\gamma \in (1, \gamma_0)$ and $M_q = c_{\gamma, \gamma_0, q} c_0 M$ for some constant $c_{\gamma, \gamma_0, q}$ depending only on (γ, γ_0, q).

5. Concentration inequalities

In this section, we consider the concentration of measure for $W_p(\hat{\mu}, \mu)$. In particular, we are interested in the following two types of concentration inequalities.

1. *Mean-concentration*. The goal is to find good upper bounds of the probability

$$\mathbb{P}\big[\big|W_p(\hat{\mu}, \mu) - \mathbb{E}W_p(\hat{\mu}, \mu)\big| > t\big]$$

for $t \in \mathbb{R}^+$.

2. 0-*concentration*. The goal is to upper bound

$$\mathbb{P}\big[W_p(\hat{\mu}, \mu) > t\big].$$

[8,16] have considered 0-concentration of $W_p^p(\hat{\mu}, \mu)$. We argue that concentration around the mean can be more informative, since in some important cases (for example, $p = 1$ and d large or infinite) the rate of concentration around the mean can be much smaller than the mean value itself. Nevertheless, a good mean-concentration can lead to a good 0-concentration, because if $\mathbb{E}W_p(\hat{\mu}, \mu) \leq R_n$ then

$$\mathbb{P}\big[W_p(\hat{\mu}, \mu) > t\big] \leq \mathbb{P}\big[W_p(\hat{\mu}, \mu) - \mathbb{E}W_p(\hat{\mu}, \mu) > (t - R_n)_+\big].$$

5.1. A Bernstein-type McDiarmid's inequality

Let X_i ($1 \leq i \leq n$) be independent (not necessarily identically distributed) samples from probability distributions μ_i on spaces $\mathcal{X}_i$, and $X_1', \ldots, X_n'$ be independent copies of each X_i. Denote $X = (X_1, \ldots, X_n)$ and $X_{(i)}' = (X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n)$, which is identical to X except that the ith entry is replaced by X_i'. Let $f : \prod_{i=1}^n \mathcal{X}_i \mapsto \mathbb{R}$ be a function such that $\mathbb{E}|f(X)| < \infty$, and define

$$D_i = f(X) - f\big(X_{(i)}'\big).$$

To establish concentration around the mean for a function of independent random inputs, a classical tool is McDiarmid's inequality, which requires the difference D_i to be bounded with probability one. This is not enough for our purpose if we take $f(X) = W_p(\hat{\mu}, \mu)$ and the support of μ is unbounded. To fix this problem, we need a stronger version of McDiarmid's inequality that holds under weaker conditions. In particular, we consider the following Bernstein-type moment condition.

$$\exists \sigma_i, M > 0 \quad \text{s.t. } \mathbb{E}\big(|D_i|^k \mid X_{-i}\big) \leq \frac{1}{2}\sigma_i^2 k! M^{k-2}, \quad \forall \text{ integer } k \geq 2, \text{ a.s.,} \tag{11}$$

where $X_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ is the random vector obtained by removing X_i from X.

When $f(X) = \sum_{i=1}^n X_i$, then equation (11) reduces to the classical Bernstein's tail condition, which is used to prove the Bernstein's inequality for sums of independent random variables. The following theorem extends the Bernstein's inequality to general functions of independent inputs.

Theorem 5.1 (Bernstein-type McDiarmid's Inequality). *Assume* (11) *holds*, *then*

$$\mathbb{P}(f - \mathbb{E}f \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2 + 2tM}\right) \tag{12}$$

where $\sigma^2 = \sum_{i=1}^n \sigma_i^2$.

Theorem 5.1 can be derived from the more general martingale concentration inequalities [10], Theorem 1.2A (see also [36]). In Appendix D, we give a simple and short proof that is similar to the standard proofs of Bernstein's inequality and McDiarmid's inequality.

Remark 3. Applying Theorem 5.1 to $-f$ leads to the same concentration bound for the event $f - \mathbb{E}f \leq -t$. Thus, Theorem 5.1 implies the absolute deviation bound

$$\mathbb{P}\big(|f - \mathbb{E}f| \geq t\big) \leq 2\exp\left(-\frac{t^2}{2\sigma^2 + 2tM}\right).$$

Remark 4. We discuss a few related works. [25] extends McDiarmid's inequality to unbounded differences by assuming sub-Gaussianity of D_i. It is well known that sub-Gaussianity is strictly stronger than Bernstein's tail condition [37]. In fact, the Bernstein's tail condition is nearly equivalent to being sub-exponential. An unpublished manuscript [44] extends McDiarmid's inequality to a Bernstein-type condition. However, this result still assumes that the differences D_i are uniformly bounded. When $\mathcal{X}_1 = \cdots = \mathcal{X}_n = \mathcal{X}$ and X_i's are i.i.d., Theorem 5.1 rediscovers a result of [11], Proposition 3.1.

5.2. Concentration of $W_p(\hat{\mu}, \mu)$

Now let $f(X) = W_p(\hat{\mu}, \mu)$, and $\hat{\mu}'$ the empirical measure given by $X'_{(i)}$. Since the ordering of sample points does not matter in $W_p(\hat{\mu}, \mu)$, the particular choice of i does not matter. By triangle inequality $|f(X) - f(X'_{(i)})| \leq W_p(\hat{\mu}, \hat{\mu}') \leq n^{-1/p}\|X_i - X'_i\|$.

If we assume that

$$\mathbb{E}\big(\|X_i\|^k\big) \le \frac{1}{2}s^2 k! V^{k-2}, \quad \forall \text{ integer } k \ge 2, \tag{13}$$

for some constants s, V, then straight calculation shows that such $f(X) = W_p(\hat{\mu}, \mu)$ satisfies the Bernstein-type tail condition (11) with $\sigma_i = 2sn^{-1/p}$ and $M = 2Vn^{-1/p}$. Thus, we have the following corollary.

Corollary 5.2. *If $X_i \sim \mu$ satisfies* (13) *with some constants s, V, then for all $t > 0$*

$$\mathbb{P}\big[\big|W_p(\hat{\mu}, \mu) - \mathbb{E}W_p(\hat{\mu}, \mu)\big| \ge t\big] \le 2\exp\left(-\frac{t^2}{8s^2 n^{1-2/p} + 4Vtn^{-1/p}}\right).$$

Corollary 5.2 is useful when $p \in [1, 2)$, implying that $W_p(\hat{\mu}, \mu) - \mathbb{E}W_p(\hat{\mu}, \mu) = O_P(n^{-(1/p-1/2)})$. In the most interesting case $p = 1$, the rate becomes the desired $n^{-1/2}$. This result is dimension-free, assuming only the tail condition (13). Condition (13) is implied by the sub-exponential tail condition. Following [37], for a random variable Z and real number $\alpha \ge 1$, define the Orlicz ψ_α norm of Z as

$$\|Z\|_{\psi_\alpha} = \inf\big\{c > 0 : \mathbb{E}e^{|Z/c|^\alpha} \le 2\big\}. \tag{14}$$

If $\|\|X_i\|\|_{\psi_1} = C < \infty$, then (13) holds with $V = C$ and $s = \sqrt{2}C$. This condition is called "sub-exponential" because $\|\|X_i\|\|_{\psi_1} < \infty$ if and only if $\mathbb{P}(\|X_i\| \ge t) \le ce^{-c't}$ for some positive constants c, c'.

Remark 5. In the special case of $p = 1$, there have been related mean-concentration results using variants of McDiarmid's inequality. For example, [6] also provides $n^{-1/2}$ mean-concentration for $W_1(\hat{\mu}, \mu)$, but requires a stronger sub-Gaussian tail of $\|X_i\|$. In a more recent work, [11] establishes a similar mean-concentration result applicable to the case $p = 1$, under a condition that can be translated to a sub-exponential tail of $\|X_i\|$ in our context.

Euclidean spaces

If $\mathcal{X} = \mathbb{R}^d$ with $d > 2$, and $\|\|X_i\|\|_{\psi_1} < \infty$, the sub-exponential condition implies finiteness of all moments, and Theorem 3.1 implies that $\mathbb{E}W_1(\hat{\mu}, \mu) \le c_1 n^{-1/d}$. Applying the one-sided version of Corollary 5.2 we get a 0-concentration bound,

$$\mathbb{P}\big[W_1(\hat{\mu}, \mu) \ge t\big] \le \mathbb{P}\big[W_1(\hat{\mu}, \mu) - \mathbb{E}W_1(\hat{\mu}, \mu) \ge t - c_1 n^{-1/d}\big]$$
$$\le \exp\big[-cn\big(t - c_1 n^{-1/d}\big)_+^2\big], \tag{15}$$

where c is a constant depending only on $\|\|X_i\|\|_{\psi_1}$. The bound given in (15) can be compared with [16], Theorem 2, equation (1), which assumes a stronger condition $\|\|X_i\|\|_{\psi_\alpha} < \infty$ for some $\alpha > 1$. In the interesting regime of $t \asymp n^{-1/d}$ (on the same scale as the optimal upper bound of $\mathbb{E}W_1(\hat{\mu}, \mu)$), the result of [16] only provides a probability bound of order $O(1)$, while (15) provides a bound of $\exp(-c(c_2 - c_1)^2 n^{1-2/d})$ for $t = c_2 n^{-1/d}$ with $c_2 > c_1$.

Functional spaces

When $\mathcal{X} = L_2$, the convergence rate of $\mathbb{E}W_1(\hat{\mu}, \mu)$ is typically sub-polynomial in n, so the polynomial concentration rate given in Corollary 5.2 when $p \in [1, 2)$ can be quite meaningful. To better understand the sub-exponential condition (14) for functional data, we consider the Karhunen–Loève decomposition (10). The sub-exponential condition can be satisfied under standard decay assumptions of the principal score variance σ_m^2, and corresponding tail bounds on the standardized principal component scores Z_m. The following proposition can be obtained after some straightforward applications of triangle inequality and some basic properties of Orlicz ψ_α norms.

Proposition 5.3. *Assuming the K-L decomposition* (10), *then* $\|\|X_i\|\|_{\psi_1} < \infty$ *if one of the following holds.*

1. $\sum_{m=1}^{\infty} \sigma_m^2 < \infty$, *and* $\sup_m \|Z_m\|_{\psi_2} < \infty$.
2. $\sum_{m=1}^{\infty} \sigma_m < \infty$, *and* $\sup_m \|Z_m\|_{\psi_1} < \infty$.

Part 1 of Proposition 5.3 covers square-integrable Gaussian processes as an important special case (i.e., $Z_m \sim N(0, 1)$ for all m). Part 2 suggests that the tail condition on the PC scores Z_m can be relaxed at the cost of a faster decay of eigenvalues.

5.3. Concentration by Lipschitz property of $W_p(\hat{\mu}, \mu)$

One can obtain concentration results for larger values of p by exploiting the Lipschitz property of $W_p(\hat{\mu}, \mu)$ as a function of $(x_1, \ldots, x_n) \in \mathcal{X}^n$. Similar ideas have been explored in the special case of $\mathcal{X} = \mathbb{R}$ (see Section 7.1 of [5], and also [19]).

For $x = (x_1, \ldots, x_n)$ and $x' = (x_1', \ldots, x_n')$, define distance $\|x - x'\|^2 = \sum_{i=1}^{n} \|x_i - x_i'\|^2$. Then one has, by triangle inequality,

$$\left| W_p(\hat{\mu}, \mu) - W_p(\hat{\mu}', \mu) \right| \le W_p(\hat{\mu}, \hat{\mu}') \le n^{-1/p} \left(\sum_{i=1}^{n} \|x_i - x_i'\|^p \right)^{1/p} \le n^{-\frac{1}{2 \vee p}} \|x - x'\|.$$

Therefore, $W_p(\hat{\mu}, \mu)$ is $n^{-\frac{1}{2 \vee p}}$-Lipschitz as a function from $\mathcal{X}^n$ to $\mathbb{R}$. Such a Lipschitz property can lead to useful mean-concentration results provided the random vector X has nice tail behavior. There are many different characterizations of such tail behavior of X [9]. Here we use log Sobolev inequality for its presentation simplicity and generality. Extension to other similar conditions, such as the Poincaré inequality, shall be straightforward.

Definition 2. A probability measure μ on a metric space $\mathcal{X}$ is said to satisfy a log Sobolev inequality with constant C, if

$$\mathbb{E}_\mu \left[f(X)^2 \log f(X)^2 \right] - \mathbb{E}_\mu \left[f(X)^2 \right] \log \mathbb{E}_\mu \left[f(X)^2 \right] \le 2C \mathbb{E}_\mu \left\| \nabla f(X) \right\|^2$$

for all smooth function $f : \mathcal{X} \mapsto \mathbb{R}$ such that the expectations are finite, where $\|\nabla f(x)\| = \sup_y \lim_{t \downarrow 0} \frac{|f(x+ty) - f(x)|}{t \|y\|}$.

The log Sobolev inequality holds for strongly log-concave densities on $\mathcal{X} = \mathbb{R}^d$. If the density of X has the form $e^{-U(x)}$ with $U \succeq I/C$, then it satisfies Definition 2 with the same constant C. For more details about log Sobolev inequalities, see [26].

Combining the Lipschitz property of $W_p(\hat{\mu}, \mu)$ and the log Sobolev inequality under product measure [26], Theorem 5.3 and Corollary 5.7, we have

Theorem 5.4. *If the distribution μ satisfies log Sobolev inequality with constant C, then*

$$\mathbb{P}\big[W_p(\hat{\mu}, \mu) - \mathbb{E}W_p(\hat{\mu}, \mu) \geq t\big] \leq \exp\left(-\frac{n^{\frac{2}{2\vee p}}t^2}{2C}\right).$$

Theorem 5.4 gives the same rate as Corollary 5.2 for $p = 1$, with strict improvement for $p > 1$ and non-trivial bounds for all values of p. In particular, with $\mathcal{X} = \mathbb{R}^d$ and a strongly log-concave distribution μ, the mean concentration of $W_p(\hat{\mu}, \mu)$ has order $n^{-1/p}$, and is dominated by the mean value $\mathbb{E}W_p(\hat{\mu}, \mu)$, which, as implied by Theorem 3.1, has order $n^{-\frac{1}{(2p)\vee d}}$.

Appendix A: Additional background for optimal transport

Definition 3 (Transport and coupling). Let μ, ν be two measures supported on the same measurable set $F \subseteq \mathcal{X}$ with the same amount of mass $\mu(F) = \nu(F) > 0$. A *coupling* between μ and ν is a measure ξ on $F \times F$ such that $\int_{y \in F} \xi(dx, dy) = \mu(dx)$ and $\int_{x \in F} \xi(dx, dy) = \nu(dy)$.

Equivalently, a coupling ξ between μ and ν can also be represented by the corresponding transition kernel $K(\cdot, \cdot)$ such that for each $x \in F$, $K(x, \cdot)$ is a measure on F, and $\int_{x \in F} K(x, \cdot)\mu(dx) = \nu$. We call such a transition kernel a *transport* from μ to ν.

The following lemma is due to [12], which is the building block for the construction of optimal transport from μ to $\hat{\mu}$.

Lemma A.1. *Let μ, ν be two measures with same mass on their common support F. Let $C_1, \ldots, C_k$ be a partition of F such that*

$$\nu|_{C_i} = \frac{\nu(C_i)}{\mu(C_i)}\mu|_{C_i}, \quad \forall 1 \leq i \leq k.$$

Then there exists a coupling ξ between μ and ν such that

$$\xi\big(\{(x, y) \in F^2 : x \neq y\}\big) = \frac{1}{2}\sum_{i=1}^{k}\big|\nu(C_i) - \mu(C_i)\big|.$$

Proof. Let $\delta = \frac{1}{2}\sum_{i=1}^{k}|\mu(C_i) - \nu(C_i)|$, and define

$$\xi = \xi_1 + \delta^{-1}\xi_2$$

where ξ_1 is the measure induced from the univariate measure $\mu \wedge \nu$ after the mapping $x \mapsto (x, x)$, and ξ_2 is the product measure $(\mu - \nu)_+ \times (\nu - \mu)_+$. The claim can be directly verified. $\square$

Appendix B: Proofs for the general bound

In this section, we provide proofs for Lemma 2.1, Lemma 2.2, and Theorem 2.3. Lemma 2.1 has been proved in [41], here we present an alternative proof based on the Markov chain method in [12]. The proofs of Lemma 2.2 and Theorem 2.3 combines the proof ideas in [16] with the general and finite partition approach.

Proof of Lemma 2.1. Let $\mu_0 = \mu$. For $\ell \geq 0$, define measure $\mu_{\ell+1}$ as

$$\mu_{\ell+1} = \sum_{F \in \mathcal{A}_{\ell+1}} \frac{\nu(F)}{\mu_\ell(F)} \mu_\ell|_F.$$

Again we use the convention $0/0 = 0$.

Facts.

1. *For all $F \in \mathcal{A}_\ell$ we have $\mu_{\ell+1}(F) = \mu_\ell(F) = \nu(F)$.*
2. $\mu_{\ell+1}|_C = \frac{\nu(C)}{\mu(C)} \mu|_C$, *for $C \in \mathcal{A}_{\ell+1}$.*
3. $W_p^p(\mu_\ell, \nu) \leq 2^p 3^{-p\ell}$.

Facts 1–2 are straightforward to verify. To see the third one, use the fact that $\mu_\ell(F) = \nu(F)$ for all $F \in \mathcal{A}_\ell$, one can construct a transport from $\mu_\ell|_F$ to $\nu|_F$ using a product measure, with cost no more than $\nu(F)[\mathrm{diam}(F)]^p$. The desired claim follows by summing over all $F \in \mathcal{A}_\ell$.

Next, we need a low-cost transport from μ_ℓ to $\mu_{\ell+1}$. Due to Facts 1 and 2 above, for each $F \in \mathcal{A}_\ell$, Lemma A.1 ensures the existence of a coupling ξ_F between $\mu_\ell|_F$ and $\mu_{\ell+1}|_F$ such that

$$\xi_F\left(\{(x, y) \in F^2 : x \neq y\}\right) \leq \frac{1}{2} \sum_{C \in \mathcal{A}_{\ell+1}, C \subseteq F} \left|\mu_{\ell+1}(C) - \mu_\ell(C)\right|$$

$$= \frac{1}{2} \sum_{C \in \mathcal{A}_{\ell+1}, C \subseteq F} \left|\nu(C) - \frac{\nu(F)}{\mu(F)} \mu(C)\right|.$$

Let

$$\xi_\ell = \sum_{F \in \mathcal{A}_\ell} \xi_F.$$

We have

$$\xi_\ell\left(\{(x, y) : x \neq y\}\right) = \frac{1}{2} \sum_{F \in \mathcal{A}_\ell} \sum_{C \in F \cap \mathcal{A}_{\ell+1}} \left|\nu(C) - \nu(F)\frac{\mu(C)}{\mu(F)}\right|$$

Let $Z_0, \ldots, Z_{\ell^*}$ be a Markov chain with $Z_0 \sim \mu$ and $(Z_{\ell+1} \mid Z_\ell = z) \sim K_\ell(z, \cdot)$, where K_ℓ is the transition kernel corresponding to ξ_ℓ. By construction, K_ℓ only moves points inside each $F \in \mathcal{A}_\ell$ so $\|Z_\ell - Z_{\ell'}\| \leq 2 \times 3^{-\ell}$ for all $\ell' > \ell$.

Let $L = \inf\{0 \leq \ell \leq \ell^* - 1 : Z_{\ell+1} \neq Z_\ell\}$. If $Z_0 = \cdots = Z_{\ell^*}$ then define $L = \infty$. By construction we have

$$Z_0 = \cdots = Z_L, \qquad \|Z_L - Z_\ell\| \leq 3^{-L}, \quad \forall \ell > L.$$

Then

$$\mathbb{E}\|Z_0 - Z_{\ell^*}\|^p$$

$$\leq \mathbb{E}\big[\mathbf{1}\big(L < \ell^*\big)\|Z_L - Z_{\ell^*}\|^p\big] \leq 2^p \mathbb{E} 3^{-pL}$$

$$= 2^p \sum_{\ell=0}^{\ell^*-1} 3^{-p\ell} \mathbb{P}(L = \ell) \leq 2^p \sum_{\ell=0}^{\ell^*-1} 3^{-p\ell} \mathbb{P}(Z_\ell \neq Z_{\ell+1})$$

$$= 2^{p-1} \sum_{\ell=0}^{\ell^*-1} 3^{-p\ell} \sum_{F \in \mathcal{A}_\ell} \sum_{C \in \mathcal{A}_{\ell+1}, C \subseteq F} \left| \nu(C) - \nu(F) \frac{\mu(C)}{\mu(F)} \right|$$

$$= 2^{p-1} \sum_{\ell=0}^{\ell^*-1} 3^{-p\ell} \sum_{F \in \mathcal{A}_\ell} \sum_{C \in \mathcal{A}_{\ell+1}, C \subseteq F} \left| \nu(C) - \mu(C) + \mu(F) \frac{\mu(C)}{\mu(F)} - \nu(F) \frac{\mu(C)}{\mu(F)} \right|$$

$$\leq 2^{p-1} \sum_{\ell=0}^{\ell^*-1} 3^{-p\ell} \sum_{F \in \mathcal{A}_\ell} \sum_{C \in \mathcal{A}_{\ell+1}, C \subseteq F} \left[\big| \nu(C) - \mu(C) \big| + \frac{\mu(C)}{\mu(F)} \big| \nu(F) - \mu(F) \big| \right]$$

$$= 2^{p-1} \sum_{\ell=0}^{\ell^*-1} 3^{-p\ell} \left[\sum_{C \in \mathcal{A}_{\ell+1}} \big| \nu(C) - \mu(C) \big| + \sum_{F \in \mathcal{A}_\ell} \big| \nu(F) - \mu(F) \big| \right]$$

$$\leq \big(1 + 3^p\big) 2^{p-1} \sum_{\ell=1}^{\ell^*} 3^{-p\ell} \sum_{F \in \mathcal{A}_\ell} \big| \nu(F) - \mu(F) \big|.$$

Let (Z_{ℓ^*}, Z) be distributed as the optimal coupling between μ_{ℓ^*} and ν. Then combining the last inequality and Fact 3 above we obtain

$$W_p^p(\mu, \nu) \leq \mathbb{E}\|Z_0 - Z\|^p \leq 2^{p-1}\big(\mathbb{E}\|Z_0 - Z_{\ell^*}\|^p + \mathbb{E}\|Z_{\ell^*} - Z\|^p\big)$$

$$\leq c_p \left[3^{-p\ell^*} + \sum_{\ell=1}^{\ell^*} 3^{-p\ell} \sum_{F \in \mathcal{A}_\ell} \big| \nu(F) - \mu(F) \big| \right]. \qquad \square$$

Proof of Lemma 2.2. The moment condition ensures that μ and ν are supported on $\bigcup_{j \geq 0} B_j$. The remaining of the proof mainly mimics that of [16].

Let $\tilde{\xi}_j$ be the optimal coupling between $\tilde{\mu}_j$ and $\tilde{\nu}_j$, and ξ_j the image of $\tilde{\xi}_j$ after the map $(x, y) \mapsto 2^j(x, y)$.

Let $\eta = \frac{1}{2}\sum_{j\geq 0}|\mu(B_j) - \nu(B_j)|$ and define

$$\xi = \sum_{j\geq 0}\left(\mu(B_j) \wedge \nu(B_j)\right)\xi_j + \eta^{-1}\alpha \times \beta$$

where

$$\alpha = \sum_{j\geq 0}\left(\mu(B_j) - \nu(B_j)\right)_+ \frac{\mu|_{B_j}}{\mu(B_j)}, \qquad \beta = \sum_{j\geq 0}\left(\nu(B_j) - \mu(B_j)\right)_+ \frac{\nu|_{B_j}}{\nu(B_j)}.$$

Then ξ is a coupling between μ and ν. It also holds that $\alpha(\mathcal{X}) + \beta(\mathcal{X}) = \eta$.

Now

$$\iint \|x - y\|^p \eta^{-1}\alpha(dx)\beta(dy)$$

$$\leq 2^{p-1}\iint \left(\|x\|^p + \|y\|^p\right)\eta^{-1}\alpha(dx)\beta(dy)$$

$$= 2^{p-1}\eta^{-1}\left[\int \|x\|^p\alpha(dx)\int \beta(dy) + \int \alpha(dx)\int \|y\|^p\beta(dy)\right]$$

$$\leq 2^{p-1}\sum_{j\geq 0}2^{jp}\left[\left(\mu(B_j) - \nu(B_j)\right)_+ + \left(\nu(B_j) - \mu(B_j)\right)_+\right]$$

$$= 2^{p-1}\sum_{j\geq 0}2^{pj}|\mu(B_j) - \nu(B_j)|.$$

Then

$$W_p^p(\mu, \nu)$$

$$\leq \iint \|x - y\|^p\xi(dx, dy)$$

$$= \sum_{j\geq 0}\left(\mu(B_j) \wedge \nu(B_j)\right)\iint \|x - y\|^p\xi_j(dx, dy) + \iint \|x - y\|^p\eta^{-1}\alpha(dx)\beta(dy)$$

$$\leq \sum_{j\geq 0}2^{jp}\left(\mu(B_j) \wedge \nu(B_j)\right)W_p^p(\tilde{\mu}_j, \tilde{\nu}_j) + 2^{p-1}\sum_{j\geq 0}2^{pj}|\mu(B_j) - \nu(B_j)|$$

$$\leq \sum_{j\geq 0}2^{pj}\left[\left(\mu(B_j) \wedge \nu(B_j)\right)W_p^p(\tilde{\mu}_j, \tilde{\nu}_j) + 2^{p-1}|\mu(B_j) - \nu(B_j)|\right], \tag{16}$$

which concludes the proof. $\qquad\square$

Proof of Theorem 2.3. Without loss of generality, we assume $\|\rho(X)\|_q = 1$, since otherwise we can always consider $X/\|\rho(X)\|_q$.

With $\nu = \hat\mu$, the plan is to use Lemma 2.1 to control the term $W_p^p(\tilde\mu_j, \tilde\nu_j)$ in Lemma 2.2. The fact that $\nu = \hat\mu$ and the moment condition imply the conditions required by Lemma 2.2. First, observe that

$$\left(\mu(B_j) \wedge \nu(B_j)\right) \sum_{\ell=1}^{\ell^*} 3^{-p\ell} \sum_{F \in \mathcal{A}_\ell} \left|\tilde\mu(F) - \tilde\nu(F)\right|$$

$$\le \mu(B_j) \sum_{\ell=1}^{\ell^*} 3^{-p\ell} \sum_{F \in \mathcal{A}_\ell} \left|\tilde\mu(F) - \tilde\nu(F)\right|$$

$$\le \mu(B_j) \sum_{\ell=1}^{\ell^*} 3^{-p\ell} \sum_{F \in \mathcal{A}_\ell} \left|\frac{\mu((2^j F) \cap B_j)}{\mu(B_j)} - \frac{\nu((2^j F) \cap B_j)}{\nu(B_j)}\right|$$

$$\le \sum_{\ell=1}^{\ell^*} 3^{-p\ell} \sum_{F \in \mathcal{A}_\ell} \left[\left|\mu\big((2^j F) \cap B_j\big) - \nu\big((2^j F) \cap B_j\big)\right| + \left|1 - \frac{\mu(B_j)}{\nu(B_j)}\right| \nu\big((2^j F) \cap B_j\big)\right]$$

$$\le \sum_{\ell=1}^{\ell^*} 3^{-p\ell} \sum_{F \in \mathcal{A}_\ell} \left|\mu\big((2^j F) \cap B_j\big) - \nu\big((2^j F) \cap B_j\big)\right| + \frac{3^{-p}}{1 - 3^{-p}}\left|\mu(B_j) - \nu(B_j)\right|.$$

Now applying Lemma 2.1 to each of $W_p^p(\tilde\mu_j, \tilde\nu_j)$, combining with the last inequality, and plugging in (16), we get

$$W_p^p(\mu, \nu)$$

$$\le c_p \sum_{j \ge 0} 2^{pj} \left[\left(\mu(B_j) \wedge \nu(B_j)\right) 3^{-p\ell^*} + \left|\mu(B_j) - \nu(B_j)\right|\right.$$

$$\left. + \sum_{\ell=1}^{\ell^*} 3^{-p\ell} \sum_{F \in \mathcal{A}_\ell} \left|\mu\big((2^j F) \cap B_j\big) - \nu\big((2^j F) \cap B_j\big)\right| + \left|\mu(B_j) - \nu(B_j)\right|\right]$$

$$\le c_p \sum_{j \ge 0} 2^{pj} \left[\mu(B_j) 3^{-p\ell^*} + \sum_{\ell=0}^{\ell^*} 3^{-p\ell} \sum_{F \in \mathcal{A}_\ell} \left|\mu\big((2^j F) \cap B_j\big) - \nu\big((2^j F) \cap B_j\big)\right|\right]. \quad (17)$$

By the moment condition, Markov's inequality implies that

$$\mu(B_j) \le 2^{-q(j-1)}.$$

Because $n\hat{\mu}(A)$ is a binomial random variable with parameters $\mu(A)$ and n, we have $\mathbb{E}|\hat{\mu}(A) - \mu(A)| \le (2\mu(A)) \wedge \sqrt{\mu(A)/n}$ for all A. Therefore,

$$\sum_{F \in \mathcal{A}_\ell} \mathbb{E}\big|\hat{\mu}\big((2^j F) \cap B_j\big) - \mu\big((2^j F) \cap B_j\big)\big|$$

$$\le \sum_{F \in \mathcal{A}_\ell} \big[2\mu\big((2^j F) \cap B_j\big)\big] \wedge \sqrt{\mu\big((2^j F) \cap B_j\big)/n}$$

$$\le \big[2\mu(B_j)\big] \wedge \left[\sqrt{\frac{\mathrm{card}(\mathcal{A}_\ell)\mu(B_j)}{n}}\right],$$

where the last step uses Cauchy–Schwarz.

Now the proof is complete if we can show that there exists a sequence of nested partitions $\mathcal{A}_1, \ldots, \mathcal{A}_{\ell^*}$ such that $\sup_{F \in \mathcal{A}} \mathrm{diam}(F) \le 2 \times 3^{-\ell}$, and $\mathrm{card}(\mathcal{A}_\ell) \le \bar{N}_\ell = N_{3^{-(\ell+1)}}(B_0)$. Proposition 3 of [41] guarantees the existence of such a sequence of partitions. $\qquad\square$

Appendix C: Proofs for Euclidean and functional spaces

Proof of Theorem 3.1. Recall that now B_0 is the unit ball in $\mathbb{R}^d$. Use the result of [39], we have, for some constant c

$$\bar{N}_\ell = N_{3^{-(\ell+1)}}(B_0) \le 3^{(\ell+c)d}$$

Then Theorem 2.3 implies

$$W_p^p(\hat{\mu}_n, \mu) \le c_p \sum_{j \ge 0} 2^{pj} \left\{ 2^{-qj} 3^{-p\ell^*} + \sum_{\ell=0}^{\ell^*} 3^{-p\ell} \big[2^{-qj} \wedge \big(3^{d(\ell+c)/2} 2^{-qj/2} n^{-1/2}\big)\big] \right\}.$$

Now let

$$\ell_j^* = \left\lfloor \frac{\log_2 n - qj}{d \log_2 3} - c \right\rfloor,$$

and

$$\ell^* = \ell_0^* = \left\lfloor \frac{\log_2 n}{d \log_2 3} - c \right\rfloor.$$

If ℓ is an integer, then

$$2^{-qj} \ge 3^{d(\ell+c)/2} 2^{-qj/2} n^{-1/2} \Leftrightarrow \ell \le \ell_j^*.$$

Let $j_n^* = \sup\{j \in \mathbb{Z} : \ell_j^* \ge 0\} = \lfloor q^{-1}(\log_2 n - cd \log_2 3) \rfloor$.

If $d \ge (2c)^{-1} \log_3 n$ then $n^{-(1/d)} \ge 3^{-2c}$ is a constant and the claim of theorem follows from Theorem 2.3, where the right-hand side is trivially bounded by a constant depending only on (p, q).

Now we focus on the case $d \geq (2c)^{-1} \log_3 n$, which implies that $j_n^* \geq 0$.
When $j \leq j_n^*$, then $\ell_j^* \geq 0$, and ("$\lesssim$" means up to a constant factor $c_{p,q}$)

$$\sum_{\ell=0}^{\ell^*} 3^{-p\ell} \left[2^{-qj} \wedge \left(3^{d(\ell+c)/2} 2^{-qj/2} n^{-1/2} \right) \right]$$

$$\leq \sum_{\ell=0}^{\ell_j^*} 3^{-p\ell} 3^{d(\ell+c)/2} 2^{-qj/2} n^{-1/2} + \sum_{\ell=\ell_j^*+1}^{\ell^*} 3^{-p\ell} 2^{-qj}$$

$$\lesssim 2^{-qj/2} n^{-1/2} 3^{cd/2} \sum_{\ell=0}^{\ell_j^*} 3^{-(p-d/2)\ell} + 2^{-qj} \left(2^{-qj} n \right)^{-p/d}.$$

When $j > j_n^*$, then $\ell_j^* < 0$, and

$$\sum_{\ell=0}^{\ell^*} 3^{-p\ell} \left[2^{-qj} \wedge \left(3^{d(\ell+c)/2} 2^{-qj/2} n^{-1/2} \right) \right] \leq \sum_{\ell=0}^{\ell^*} 3^{-p\ell} 2^{-qj} \lesssim 2^{-qj}.$$

Thus, $\sum_{j > j_n^*} 2^{pj} 2^{-qj} \lesssim 2^{-qj_n^*(1-p/q)} \lesssim (n3^{-cd})^{-(1-p/q)}$ and

$$\sum_{j \geq 0} 2^{pj} \sum_{\ell=0}^{\ell^*} 3^{-p\ell} \left[2^{-qj} \wedge \left(3^{d(\ell+c)/2} 2^{-qj/2} n^{-1/2} \right) \right]$$

$$\lesssim \sum_{j=0}^{j_n^*} 2^{pj} \left\{ 2^{-qj/2} n^{-1/2} 3^{cd/2} \sum_{\ell=0}^{\ell_j^*} 3^{-(p-d/2)\ell} + 2^{-qj} \left(2^{-qj} n \right)^{-p/d} \right\}$$

$$+ \left(n3^{-cd} \right)^{-(1-p/q)}. \tag{18}$$

Case 1: $p > d/2$. The terms in the sum $\sum_{\ell=0}^{\ell_j^*} 3^{-(p-d/2)\ell}$ is geometrically decreasing and hence the sum is bounded by the first term by a constant factor. Moreover, for $j \leq j_n^*$ we have $(2^{-qj} n)^{-p/d} \leq c_{p,q} (2^{-qj} n)^{-1/2}$. So (18) is bounded by (ignoring constant factors)

$$\sum_{j=0}^{j_n^*} 2^{pj} 2^{-qj/2} n^{-1/2} + n^{-(1-p/q)}. \tag{19}$$

Case 1.1: $p < q/2$. Both terms in (19) are bounded by $n^{-1/2}$.

Case 1.2: $p = q/2$. The first term in (19) equals $(1 + j_n^*) n^{-1/2} \leq c_{p,q} (\log n) n^{-1/2}$, and the second term in (19) equals $n^{-1/2}$.

Case 1.3: $p \in (q/2, q)$. The first sum in (19) equals $n^{-1/2} 2^{(p-q/2)j_n^*} \leq c_{p,q} n^{-(1-p/q)}$, the same as the second term.

Case 2: $p = d/2$. Now (18) reduces to (ignoring constant factors)

$$\sum_{j=0}^{j_n^*} 2^{pj} 2^{-qj/2} n^{-1/2} \left(\ell_j^* + 1\right) + n^{-(1-p/q)}. \tag{20}$$

Case 2.1: $p < q/2$. The first term is a geometric sum bounded by $n^{-1/2} \log n$, which dominates the second term.

Case 2.2: $p = q/2$. The first term is of order $n^{-1/2} (\log n)^2$, which dominates the second term.

Case 2.3: $p > q/2$. The first term becomes $n^{-1/2} \sum_{j=0}^{j_n^*} 2^{(p-q/2)j} (\ell_j^* + 1)$. Let $\tilde{j}_n^* = q^{-1}(\log_2 n - cd \log_2 3)$, and $a = 2^{p-q/2}$

$$\sum_{j=0}^{j_n^*} 2^{(p-q/2)j} \left(\ell_j^* + 1\right) \leq \sum_{j=0}^{j_n^*} 2^{(p-q/2)j} \left(\frac{\log_2 n - qj}{d \log_2 3} - c + 1\right)$$

$$= \frac{q}{d \log_2 3} \sum_{j=0}^{j_n^*} a^j \left(\tilde{j}_n^* - j + c_{p,q}\right)$$

$$\leq c_{p,q} \sum_{j=0}^{j_n^*} a^j \left(j_n^* - j\right) + c_{p,q} a^{j_n^*}. \tag{21}$$

Standard calculation shows that

$$\sum_{j=0}^{j_n^*} a^j \left(j_n^* - j\right) = \frac{a}{(a-1)^2} \left(a^{j_n^*} - 1\right) - \frac{j_n^*}{a-1}$$

$$\leq c_{p,q} a^{j_n^*}.$$

Thus (21) is bounded by

$$c_{p,q} 2^{(p-q/2)j_n^*} \leq c_{p,q} n^{p/q - 1/2}$$

and (20) is bounded by $c_{p,q} n^{-(1-p/q)}$.

Case 3: $p < d/2$. In this case $3^{-(p-d/2)\ell}$ is an increasing geometric sequence as ℓ changes from 0 to ℓ_j^*. Thus (18) becomes, using the fact that $3^{\ell_j^*} \leq 3^{-c} (n 2^{-qj})^{1/d}$,

$$\sum_{j=0}^{j_n^*} 2^{pj} \left\{ 2^{-qj/2} n^{-1/2} 3^{cd/2} 3^{(d/2-p)\ell_j^*} + 2^{-qj} \left(2^{-qj} n\right)^{-p/d} \right\}$$

$$+ \left(n 3^{-cd}\right)^{-(1-p/q)}$$

$$\leq \sum_{j=0}^{j_n^*} 2^{-jq(1-p/d-p/q)} n^{-p/d} + \left(n 3^{-cd}\right)^{-(1-p/q)}. \tag{22}$$

Case 3.1: $d < \frac{qp}{q-p}$ (or $1 - p/d - p/q < 0$). The first term is the sum of an increasing geometric sequence, which is bounded by the last term $n^{-p/d} 2^{j_n^* q(p/d+p/q-1)} \le c_{p,q} n^{-(1-p/q)}$, matching the second term.

Case 3.2: $d = \frac{qp}{q-p}$ (or $1 - p/d - p/q = 0$). The first term is of order $n^{-p/d} \log n$ and dominates the second term.

Case 3.3: $d > \frac{qp}{q-p}$ (or $1 - p/d - p/q > 0$). The first term is of order $n^{-p/d}$. Now if $d \le 2qp/(q-p)$ then the second term is bounded by $n^{-(1-p/q)}$ and hence dominated by the first term. If $d > 2qp/(q-p)$, then by assumption that $d \le (2c)^{-1} \log_3 n$ we have $3^{-cd} \ge n^{-1/2}$ so the second term is bounded by $n^{-(1-p/q)/2}$ which is dominated by the first term.

Finally, note that the contribution from $\sum_{j \ge 0} 2^{-pj} 2^{-qj} 3^{-p\ell^*}$ is bounded by $n^{-p/d}$, which is dominated in all the nine cases. $\square$

Proof of Theorem 4.1. Again it suffices to prove for the case $M_q = 1$. In our proof "$\lesssim$" means inequality holds up to a constant depending only on (p, q, b).

According to Corollary 2.4 of [29], we have, as $\epsilon \to 0$,

$$\log N_\epsilon(B_0) = (1 + o(1)) b \epsilon^{-\frac{1}{b}}.$$

As a consequence we have, for some positive constant c_b fixed through this proof, and all $\ell \in \mathbb{N}$,

$$\bar{N}_\ell \le 2^{c_b 3^{\ell/b}}.$$

Again, in order to use Theorem 2.3, the key is to break down the term

$$2^{-qj} \wedge \left(\bar{N}_\ell^{1/2} 2^{-qj/2} n^{-1/2} \right).$$

Let

$$\ell_j^* = \left\lfloor b \left[\log_3(\log_2 n - qj) - \log_3 c_b \right] \right\rfloor$$

and $\ell^* = \ell_0^* = \lfloor b(\log_3 \log_2 n - \log_3 c_b) \rfloor$ so that $\ell \le \ell_j^* \Rightarrow 2^{-qj} \ge \bar{N}_\ell^{1/2} 2^{-qj/2} n^{-1/2}$.

Let $j_n^* = \lfloor q^{-1}(\log_2 n - c_b) \rfloor$ so that $j \le j_n \Leftrightarrow \ell_j^* \ge 0$.

If $\log_2 n < c_b$ then $n < 2^{c_b} = O(1)$. The claim of the theorem follows trivially from Theorem 2.3. Now we focus on the case $\log_2 n \ge c_b$ so that $j_n^* \ge 0$.

So when $j \le j_n^*$,

$$\sum_{\ell=0}^{\ell^*} 3^{-p\ell} \left[2^{-qj} \wedge \left(\bar{N}_\ell^{1/2} 2^{-qj/2} n^{-1/2} \right) \right]$$

$$= 2^{-qj/2} n^{-1/2} \sum_{\ell=0}^{\ell_j^*} 3^{-p\ell} 2^{(c_b/2) 3^{\ell/b}} + 2^{-qj} \sum_{\ell=\ell_j^*+1}^{\ell^*} 3^{-p\ell}$$

$$\lesssim 2^{-qj/2} n^{-1/2} + 2^{-qj} 3^{-p\ell_j^*}$$

$$\lesssim 2^{-qj/2} n^{-1/2} + 2^{-qj} (\log_2 n - qj)^{-bp}.$$

To see the first inequality, observe that the terms $3^{-p\ell}2^{(c_b/2)3^{\ell/b}}$ becomes super-geometrically increasing with rate at least 2 after a constant number of terms. Thus, the sum is bounded by a constant plus the last term, which is $3^{-p\ell_j^*}2^{(c_b/2)3^{\ell_j^*/b}} \le 3^{-p\ell_j^*}n^{1/2}2^{-qj/2}$.

When $j > j_n^*$, we have $\ell_j^* < 0$ and hence

$$\sum_{\ell=0}^{\ell^*}3^{-p\ell}\big[2^{-qj} \wedge \big(\bar{N}_\ell^{1/2}2^{-qj/2}n^{-1/2}\big)\big] \lesssim 2^{-qj}.$$

Combining the above two inequalities with Theorem 2.3 we obtain

$$\mathbb{E}W_p^p(\hat{\mu},\mu)$$

$$\lesssim 3^{-p\ell^*} + \sum_{j=0}^{j_n^*}2^{pj}\big[2^{-qj/2}n^{-1/2} + 2^{-qj}(\log_2 n - qj)^{-bp}\big] + \sum_{j>j_n^*}2^{pj}2^{-qj}$$

$$\lesssim \sum_{j=0}^{j_n^*}2^{-(q-p)j}(\log_2 n - qj)^{-bp} + n^{-1/2}\sum_{j=0}^{j_n^*}2^{-(q/2-p)j}$$

$$+ (\log n)^{-bp} + n^{-(1-p/q)}, \tag{23}$$

where the $(\log n)^{-bp}$ upper bounds $3^{-p\ell^*}$, and $n^{-(1-p/q)}$ upper bounds $\sum_{j>j_n^*}2^{pj}2^{-qj}$.

Now we control the first two terms in (23). For the first term, if we pick c_b large enough so that $c_b \ge \frac{q2^{(q/p-1)/(2b)}}{2^{(q/p-1)/(2b)}-1}$, then for all $0 \le j \le j_n^*$ the terms $2^{-(q-p)j}(\log_2 n - qj)^{-bp}$ in the first sum decreases super-geometrically with rate $2^{-(q-p)/2}$ and hence bounded by the first term, which is of order $(\log n)^{-pb}$.

For the second term, we consider three cases.

Case 1: $p < q/2$, then the sequence $2^{-(q/2-p)j}$ is geometrically decreasing and the sum is bounded by the first term which is 1. So the second part in (23) is bounded by $n^{-1/2}$.

Case 2: $p = q/2$, the sum becomes $n^{-1/2}(1 + j_n^*) \lesssim n^{-1/2}\log n$.

Case 3: $p > q/2$, the sum is dominated by the last term, which is $n^{-1/2}2^{(p-q/2)j_n^*} \lesssim n^{-1/2}n^{p/q-1/2} = n^{-(1-p/q)}$.

In all these cases, (23) is dominated by the first term and hence

$$\mathbb{E}W_p^p(\hat{\mu},\mu) \lesssim (\log n)^{-bp},$$

which implies the desired upper bound using $p \ge 1$ and Jensen's inequality.

Now we prove the lower bound. Let $\epsilon_n = (\frac{2}{b}\log n)^{-b}$. By Corollary 2.4 of [29], we have

$$\lim_{n\to\infty}\frac{\log N_{\epsilon_n}(B_0)}{2\log n} = 1.$$

Therefore, there exists a constant $n_0 \in \mathbb{N}$ such that $N_{\epsilon_n} \ge n$ for all $n \ge n_0$.

For $n < n_0$, the result is trivial because the desired lower bound is essentially a constant and one can pick μ_n to be a distribution with two point masses that are a constant distance apart.

For $n \geq n_0$, let $x_1, \ldots, x_n$ be points in B_0 such that $\min_{1 \leq i < j \leq n} \|x_i - x_j\| \geq \epsilon_n$. Let μ_n be the measure that puts n^{-1} mass at each x_i ($1 \leq i \leq n$).

If $\hat{\mu}_n(\{x_i\}) = 0$ for some x_i, then any transport between $\hat{\mu}_n$ and μ_n costs probability mass at least n^{-1} and distance at least ϵ_n to just cover the point x_i alone. Thus, we have the following lower bound

$$W_p(\hat{\mu}_n, \mu_n) \geq \left\{ \sum_{i=1}^{n} \epsilon_n^p n^{-1} \mathbf{1}\big[\hat{\mu}_n(\{x_i = 0\})\big] \right\}^{1/p} = \epsilon_n \kappa_n^{1/p},$$

where κ_n is the proportion of x_i's such that $\hat{\mu}_n(\{x_i\}) = 0$. As a result

$$\mathbb{E} W_p(\hat{\mu}_n, \mu_n) \geq \epsilon_n \mathbb{E}\big(\kappa_n^{1/p}\big).$$

Elementary calculation shows that $\mathbb{E}\kappa_n \to e^{-1}$. Also κ_n has bounded difference in the sense that if one out of the n independent sample points is changed arbitrarily, then κ_n changes no more than $1/n$. Thus, by McDiarmid's inequality, $\mathbb{P}(|\kappa_n - \mathbb{E}\kappa_n| \geq t) \leq 2\exp(-2nt^2)$ for all $t > 0$. Now for n larger than some constant n_1 we have $\mathbb{E}\kappa_n \geq (2e)^{-1}$. Choosing $t = (4e)^{-1}$ we have $\mathbb{P}(\kappa_n \geq (4e)^{-1}) \to 1$. Therefore $\mathbb{E}(\kappa_n^{1/p}) \geq c$ for some universal positive constant c.   □

Proof of Lemma 4.3. For $\epsilon \in (0,1)$, let $J_1 = \lceil \log_\gamma(\epsilon^{-1}) \rceil$, so that $\tau_m/\epsilon > 1 \Leftrightarrow m \leq J_1$ Let $B_1 = \{x \in \mathbb{R}^{J_1} : \sum_{m=1}^{J_1}(x_m/\tau_m)^2 \leq 1\}$. Then B_1 is the projection of B_0 onto the first J_1 coordinates. Thus $N_\epsilon(B_0) \geq N_\epsilon(B_1)$. But Theorem 1 of [14] implies that

$$\log N_\epsilon(B_1) \geq \sum_{m=1}^{J_1} \log(\tau_m/\epsilon) = J_1 \log(\epsilon^{-1}) - \log\gamma \frac{J_1(J_1 - 1)}{2}$$

and the lower bound claim follows by $\log_\gamma(\epsilon^{-1}) \leq J_1 < \log_\gamma(\epsilon^{-1}) + 1$.

For the upper bound, let $\theta \in (0, 1/2)$ be a constant. For example, we can pick $\theta = 1/3$. Define $J_2 = \lceil \log_\gamma(\epsilon\sqrt{1-\theta})^{-1} \rceil$, so that $\tau_m/\epsilon \in (\sqrt{1-\theta}, 1] \Leftrightarrow m \in (J_1, J_2]$. Let

$$B_2 = \left\{ x \in \mathbb{R}^{J_2} : \sum_{m=1}^{J_2}(x_m/\tau_m)^2 \leq 1 \right\},$$

$$B_3 = \left\{ x \in \mathbb{R}^\infty : \sum_{m \geq 1}(x_m/\tau_{J_2+m})^2 \leq 1 \right\}.$$

Then $B_0 \subset B_2 \times B_3$.

By construction $\tau_{J_2+1}/\epsilon \leq \sqrt{1-\theta}$, so B_3 can be covered by the ball centered at 0 with radius $\sqrt{1-\theta}\epsilon$. As a consequence,

$$N_{\sqrt{2-\theta}\epsilon}(B_0) \leq N_{\sqrt{2-\theta}\epsilon}(B_2 \times B_3) \leq N_\epsilon(B_2).$$

By Theorem 2 of [14], we have (c_0, c_1, depending only on γ, θ, may change from line to line)

$$\log N_\epsilon(B_2) \le \sum_{m=1}^{J_1} \log(\tau_m/\epsilon) + J_2 \log(3/\theta)$$

$$= \frac{\log \gamma}{2} \lceil \log_\gamma(\epsilon^{-1}) \rceil (2\log_\gamma(\epsilon^{-1}) - \lceil \log_\gamma(\epsilon^{-1}) \rceil + 1)$$

$$+ \left\lceil \log_\gamma(\epsilon^{-1}) + \log_\gamma \frac{1}{\sqrt{1-\theta}} \right\rceil \log(3/\theta)$$

$$\le \frac{\log \gamma}{2} \left(\log_\gamma(\epsilon^{-1}) + 1 \right)^2 + c_1 \log_\gamma(\epsilon^{-1}) + c_0$$

$$= \frac{1}{2 \log \gamma} \left(\log(\epsilon^{-1}) + c_1 \right)^2 + c_0.$$

The desired result can be obtained by plugging in $\epsilon = 3^{-(\ell+1)}/\sqrt{2-\theta}$. □

Proof of Theorem 4.2. By Lemma 4.3, we have for $\ell \in \mathbb{N}$

$$\log \bar{N}_\ell \le c_\gamma(\ell + c_1)^2 + \log c_0$$

for $c_\gamma = \frac{(\log 3)^2}{2\log \gamma}$ and positive constants c_1, c_0 depending on γ only. Then

$$\bar{N}_\ell \le c_0 e^{c_\gamma(\ell+c_1)^2}.$$

Let $j_n^* = \lfloor q^{-1} \frac{\log(n/c_0) - c_1^2 c_\gamma}{\log 2} \rfloor$.

If $j_n^* < 0$, then $n < c_0 e^{c_1^2 c_\gamma}$ and the claim follows trivially. Now we focus on the case $j_n^* \ge 0$.
Let $\ell_j^* = \lfloor [\frac{(\log(n/c_0) - qj \log 2)_+}{c_\gamma}]^{1/2} - c_1 \rfloor$, and $\ell^* = \ell_0^* = \lfloor [\frac{\log(n/c_0)}{c_\gamma}]^{1/2} - c_1 \rfloor$.
When $j \le j_n^*$, we have $\ell_j^* \ge 0$ so

$$\sum_{\ell=0}^{\ell^*} 3^{-p\ell} [2^{-qj} \wedge (\bar{N}_\ell^{1/2} 2^{-qj/2} n^{-1/2})]$$

$$\lesssim c_0^{1/2} 2^{-qj/2} n^{-1/2} \sum_{\ell=0}^{\ell_j^*} 3^{-p\ell} e^{c_\gamma(\ell+c_1)^2/2} + 2^{-qj} \sum_{\ell=\ell_j^*+1}^{\ell^*} 3^{-p\ell}$$

$$\lesssim 2^{-qj/2} n^{-1/2} + 2^{-qj} 3^{-p\sqrt{\frac{\log(n/c_0) - qj \log 2}{c_\gamma}}}.$$

To see the last inequality, observe that after a constant number of terms in the first sum in the line above, the terms $3^{-p\ell} e^{c_\gamma(\ell+c_1)^2/2}$ become super-geometrically increasing with rate at least 2. So the first sum is upper bounded (up to constant factor) by the last term plus 1.

When $j > j_n^*$, we have $\ell_j^* = 0$ and

$$\sum_{\ell=0}^{\ell^*} 3^{-p\ell}\big[2^{-qj} \wedge (\bar{N}_\ell^{1/2} 2^{-qj/2} n^{-1/2})\big] \lesssim 2^{-qj}.$$

Combining the above two inequalities with Theorem 2.3, we get

$$\mathbb{E}W_p^p(\hat{\mu}, \mu)$$

$$\lesssim 3^{-p\ell^*} + \sum_{j=0}^{j_n^*} 2^{pj} 2^{-qj/2} n^{-1/2} + \sum_{j=0}^{j_n^*} 2^{pj} 2^{-qj} 3^{-p\sqrt{\frac{\log(n/c_0)-qj\log 2}{c_\gamma}}} + \sum_{j>j_n^*} 2^{pj} 2^{-qj}$$

$$\lesssim \sum_{j=0}^{j_n^*} 2^{pj} 2^{-qj} 3^{-p\sqrt{\frac{\log(n/c_0)-qj\log 2}{c_\gamma}}} + \sum_{j=0}^{j_n^*} 2^{pj} 2^{-qj/2} n^{-1/2}$$

$$+ 3^{-p\sqrt{\log n/c_\gamma}} + n^{-(1-p/q)}, \tag{24}$$

where the term $3^{-p\sqrt{\log n/c_\gamma}}$ controls $3^{-p\ell^*}$, and $3^{-p\sqrt{\log n/c_\gamma}} + n^{-(1-p/q)}$ controls $\sum_{j>j_n^*} 2^{pj} \times 2^{-qj}$.

If we choose c_1 large enough so that $c_1 \geq \frac{qp}{q-p} \frac{\log 3}{c_\gamma}$ then it can be verified that the terms in the first sum in (24) is super-geometrically decreasing with rate $2^{-(q-p)}$. Thus, the sum is bounded by the first term $3^{-p\sqrt{(\log(n/c_0))/c_\gamma}} \lesssim 3^{-p\sqrt{(\log n)/c_\gamma}}$.

The second sum in (24) can be controlled using the same argument as for the second term in (23) in the proof of Theorem 4.1, which is bounded by $n^{-[(1-p/q)\wedge(1/2)]}(\log n)^{\mathbf{1}(q=2p)}$, and is dominated by the first term in (24).

Therefore, the first and third terms dominate in (24) and final rate is

$$\mathbb{E}W_p^p(\hat{\mu}, \hat{\mu}) \lesssim 3^{-p\sqrt{(\log n)/c_\gamma}} = e^{-p\sqrt{2\log\lambda\log n}},$$

which concludes the proof of upper bound.

For the lower bound. Let $\epsilon_n = e^{-\sqrt{2\log\gamma\log n}}$ then by Lemma 4.3 we have

$$\log N_\epsilon \geq \frac{1}{2\log\gamma}\big[\log(1/\epsilon)\big]^2 = \log n.$$

So $N_\epsilon \geq n$. Let $x_1, \ldots, x_n \in B_0$ be such that $\min_{i\neq j} \|x_i - x_j\| \geq \epsilon$. Let μ_n be the distribution putting n^{-1} mass at each x_i. The rest of the proof are identical to the lower bound proof of Theorem 4.1. $\qquad\square$

Proof of Proposition 4.4. We only prove for the polynomial decay case, the exponential case is similar. Without loss of generality, we assume $c_0 = M = 1$, as both constants can be recovered

after scaling. Then

$$\|\rho_\tau(X)\|_q^q = \mathbb{E}\left\{\left[\sum_{m\geq 1}\left(\frac{X_m}{\tau_m}\right)^2\right]^{q/2}\right\} = \left\|\sum_{m\geq 1}\left(\frac{X_m}{\tau_m}\right)^2\right\|_{q/2}^{q/2}$$

$$\leq \left[\sum_{m\geq 1}\left\|\left(\frac{X_m}{\tau_m}\right)^2\right\|_{q/2}\right]^{q/2} = \left[\sum_{m\geq 1}\left\|\frac{X_m}{\tau_m}\right\|_q^2\right]^{q/2}$$

$$= \left[\sum_{m\geq 1}\frac{\|X_m\|_q^2}{\tau_m^2}\right]^{q/2} \leq \left[\sum_{m\geq 1}(\sigma_m/\tau_m)^2\right]^{q/2},$$

where the inequality uses Minkowski inequality since we assume $q/2 \geq 1$. The claimed results follow by replacing σ_m by the assumed upper bound, and τ_m by the assumed form in (6) or (8). $\square$

Appendix D: Proof of concentration inequality

Proof of Theorem 5.1. Let $\Delta_i = \mathbb{E}[f(X_1,\ldots,X_n)\mid X_1^i] - \mathbb{E}[f(X_1,\ldots,X_n)\mid X_1^{i-1}]$, where X_i^j denotes $(X_i, X_{i+1},\ldots,X_j)$.

For two random vectors (Y, Z) and a function g, we use the notation $\mathbb{E}_Y g(Y, Z)$ to denote the conditional expectation of $g(Y, Z)$ given Z (the notation means that we integrate over Y in $g(Y, Z)$).

For any i and $c > 0$

$$\mathbb{E}\left(e^{c\Delta_i}\mid X_1^{i-1}\right) = \mathbb{E}_{X_i}\exp\left\{c\left[\mathbb{E}_{X_{i+1}^n}f(X) - \mathbb{E}_{X_{i+1}^n,X_i'}f\left(X_{(i)}'\right)\right]\right\}$$

$$= \mathbb{E}_{X_i}\exp\left\{\mathbb{E}_{X_{i+1}^n,X_i'}c\left[f(X) - f\left(X_{(i)}'\right)\right]\right\}$$

$$\leq \mathbb{E}_{X_i}\mathbb{E}_{X_{i+1}^n,X_i'}\exp(cD_i)$$

$$= \mathbb{E}_{X_{i+1}^n}\mathbb{E}_{X_i,X_i'}\exp(cD_i).$$

Using the fact that $\mathbb{E}_{X_i,X_i'}D_i = 0$, we get

$$\mathbb{E}_{X_i,X_i'}\exp(cD_i) = \mathbb{E}_{X_i,X_i'}\left(1 + cD_i + \sum_{k\geq 2}\frac{c^k D_i^k}{k!}\right)$$

$$= 1 + \mathbb{E}_{X_i,X_i'}\sum_{k\geq 2}\frac{c^k D_i^k}{k!}$$

$$\leq 1 + \frac{1}{2}\sigma_i^2 c^2 \sum_{k\geq 0}(cM)^k$$

$$\leq \exp\left(\frac{1}{2}\sigma_i^2 c^2 \sum_{k\geq 0}(cM)^k\right)$$

$$= \exp\left(\frac{1}{2}\sigma_i^2 c^2 (1-cM)^{-1}\right),$$

provided that $cM < 1$.

For any $t, c > 0$, $c < M^{-1}$, we have

$$\mathbb{P}(f - \mathbb{E}f \geq t\sigma) \leq e^{-ct\sigma}\mathbb{E}e^{c(f-\mathbb{E}f)} = e^{-ct\sigma}\mathbb{E}e^{c\sum_{i=1}^{n}\Delta_i}$$

$$\leq e^{-ct\sigma}\mathbb{E}\left[e^{c\sum_{i=1}^{n-1}\Delta_i}\mathbb{E}\left(e^{c\Delta_n} \mid X_1^{n-1}\right)\right]$$

$$\leq e^{-ct\sigma}\exp\left(\frac{1}{2}\sigma_n^2 c^2 (1-cM)^{-1}\right)\mathbb{E}\left[e^{c\sum_{i=1}^{n-1}\Delta_i}\right]$$

$$\leq e^{-ct\sigma}\prod_{i=1}^{n}\exp\left(\frac{1}{2}\sigma_i^2 c^2 (1-cM)^{-1}\right)$$

$$= e^{-ct\sigma}\exp\left(\frac{1}{2}\sigma^2 c^2 (1-cM)^{-1}\right),$$

where $\sigma^2 = \sum_{i=1}^{n}\sigma_i^2$.

The claimed result follows by choosing $c = \frac{t}{\sigma + tM}$. □

Acknowledgements

References

[1] Airoldi, E.M., Costa, T.B. and Chan, S.H. (2013). Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems* 692–700.

[2] Aldous, D.J. (1981). Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.* **11** 581–598. MR0637937 https://doi.org/10.1016/0047-259X(81)90099-3

[3] Arjovsky, M., Chintala, S. and Bottou, L. (2017). Wasserstein GAN. Available at arXiv:1701.07875.

[4] Bickel, P.J. and Chen, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.

[5] Bobkov, S. and Ledoux, M. (2014). One-dimensional empirical measures, order statistics and Kantorovich transport distances. Preprint.

[6] Boissard, E. (2011). Simple bounds for convergence of empirical and occupation measures in 1-Wasserstein distance. *Electron. J. Probab.* **16** 2296–2333. MR2861675 https://doi.org/10.1214/EJP.v16-958

[7] Boissard, E. and Le Gouic, T. (2014). On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Ann. Inst. Henri Poincaré Probab. Stat.* **50** 539–563. MR3189084 https://doi.org/10.1214/12-AIHP517

[8] Bolley, F., Guillin, A. and Villani, C. (2007). Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probab. Theory Related Fields* **137** 541–593. MR2280433 https://doi.org/10.1007/s00440-006-0004-7

[9] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities*: *A Nonasymptotic Theory of Independence*. Oxford: Oxford Univ. Press. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

[10] de la Peña, V.H. (1999). A general class of exponential inequalities for martingales and ratios. *Ann. Probab.* **27** 537–564. MR1681153 https://doi.org/10.1214/aop/1022677271

[11] Dedecker, J. and Fan, X. (2015). Deviation inequalities for separately Lipschitz functionals of iterated random functions. *Stochastic Process*. *Appl.* **125** 60–90. MR3274692 https://doi.org/10.1016/j.spa.2014.08.001

[12] Dereich, S., Scheutzow, M. and Schottstedt, R. (2013). Constructive quantization: Approximation by empirical measures. *Ann. Inst. Henri Poincaré Probab. Stat.* **49** 1183–1203. MR3127919 https://doi.org/10.1214/12-AIHP489

[13] Dudley, R.M. (1968). The speed of mean Glivenko–Cantelli convergence. *Ann. Math. Stat.* **40** 40–50. MR0236977 https://doi.org/10.1214/aoms/1177697802

[14] Dumer, I. (2006). Covering an ellipsoid with equal balls. *J. Combin. Theory Ser. A* **113** 1667–1676. MR2269546 https://doi.org/10.1016/j.jcta.2006.03.021

[15] Dumer, I., Pinsker, M.S. and Prelov, V.V. (2004). On coverings of ellipsoids in Euclidean spaces. *IEEE Trans*. *Inform*. *Theory* **50** 2348–2356. MR2097051 https://doi.org/10.1109/TIT.2004.834759

[16] Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields* **162** 707–738. MR3383341 https://doi.org/10.1007/s00440-014-0583-7

[17] Gao, C., Lu, Y. and Zhou, H.H. (2015). Rate-optimal graphon estimation. *Ann*. *Statist*. **43** 2624–2652. MR3405606 https://doi.org/10.1214/15-AOS1354

[18] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* 2672–2680.

[19] Gozlan, N. and Léonard, C. (2010). Transport inequalities. A survey. *Markov Process*. *Related Fields* **16** 635–736. MR2895086

[20] Hall, P., Müller, H.-G. and Wang, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann*. *Statist*. **34** 1493–1517. MR2278365 https://doi.org/10.1214/009053606000000272

[21] Hoover, D.N. (1982). Row-column exchangeability and a generalized model for probability. In *Exchangeability in Probability and Statistics* (*Rome*, 1981) 281–291. Amsterdam: North-Holland. MR0675982

[22] Kallenberg, O. (1989). On the representation theorem for exchangeable arrays. *J. Multivariate Anal*. **30** 137–154. MR1003713 https://doi.org/10.1016/0047-259X(89)90092-4

[23] Kantorovič, L.V. and Rubinšteĭn, G.Š. (1958). On a space of completely additive functions. *Vestnik Leningrad Univ. Math.* **13** 52–59. MR0102006

[24] Kantorovitch, L. (1942). On the translocation of masses. *C. R.* (*Dokl.*) *Acad. Sci. URSS* **37** 199–201. MR0009619

[25] Kontorovich, A. (2014). Concentration in unbounded metric spaces and algorithmic stability. In *International Conference on Machine Learning* 28–36.

[26] Ledoux, M. (1999). Concentration of measure and logarithmic Sobolev inequalities. In *Séminaire de Probabilités*, *XXXIII*. *Lecture Notes in Math*. **1709** 120–216. Berlin: Springer. MR1767995 https://doi.org/10.1007/BFb0096511

[27] Lei, J. (2018). Network representation using graph root distributions. Available at arXiv:1802.09684.

[28] Lovász, L. (2012). *Large Networks and Graph Limits*. *American Mathematical Society Colloquium Publications* **60**. Providence, RI: Amer. Math. Soc. MR3012035 https://doi.org/10.1090/coll/060

[29] Luschgy, H. and Pagès, G. (2004). Sharp asymptotics of the Kolmogorov entropy for Gaussian measures. *J. Funct. Anal*. **212** 89–120. MR2065239 https://doi.org/10.1016/j.jfa.2003.09.004

[30] Monge, G. (1781). Mémoire sur la théorie des déblais des remblais. History of the Royal Academy of Sciences of Paris.

[31] Ramsay, J.O. (2006). *Functional Data Analysis*. New York: Wiley Online Library.

[32] Singh, S. and Póczos, B. (2018). Minimax distribution estimation in Wasserstein distance. Available at arXiv:1802.08855.

[33] Sommerfeld, M. and Munk, A. (2018). Inference for empirical Wasserstein distances on finite spaces. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **80** 219–238. MR3744719 https://doi.org/10.1111/rssb.12236

[34] Talagrand, M. (1994). The transportation cost from the uniform measure to the empirical measure in dimension ≥ 3. *Ann. Probab*. **22** 919–959. MR1288137

[35] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. New York: Springer. MR2724359 https://doi.org/10.1007/b13794

[36] van de Geer, S. (1995). Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Ann. Statist*. **23** 1779–1801. MR1370307 https://doi.org/10.1214/aos/1176324323

[37] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*. *Springer Series in Statistics*. New York: Springer. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

[38] Vasershtein, L.N. (1969). Markov processes over denumerable products of spaces describing large system of automata. *Problemy Peredachi Informatsii* **5** 64–72. MR0314115 https://doi.org/10.1016/s0016-0032(33)90010-1

[39] Verger-Gaugry, J.-L. (2005). Covering a ball with smaller equal balls in $\mathbb{R}^n$. *Discrete Comput. Geom*. **33** 143–155. MR2105755 https://doi.org/10.1007/s00454-004-2916-2

[40] Villani, C. (2009). *Optimal Transport*: *Old and New*. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **338**. Berlin: Springer. MR2459454 https://doi.org/10.1007/978-3-540-71050-9

[41] Weed, J. and Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli* **25** 2620–2648. MR4003560 https://doi.org/10.3150/18-BEJ1065

[42] Williamson, R.C., Smola, A.J. and Schölkopf, B. (2001). Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Trans. Inform. Theory* **47** 2516–2532. MR1873936 https://doi.org/10.1109/18.945262

[43] Wolfe, P.J. and Olhede, S.C. (2013). Nonparametric graphon estimation. Available at arXiv:1309.5936.

[44] Ying, Y. McDiarmid's inequalities of Bernstein and Bennett forms. Available at https://www.albany.edu/~yy298919/McDiarmid.pdf.