QUARTERLY OF APPLIED MATHEMATICS
VOLUME , NUMBER 0
XXXX XXXX, PAGES 000–000

ON THE RATE OF CONVERGENCE OF EMPIRICAL MEASURE IN $\infty-$WASSERSTEIN DISTANCE FOR UNBOUNDED DENSITY FUNCTION

BY

ANNING LIU (*Department of Mathematical Sciences, Tsinghua University, Beijing 100084, People's Republic of China*),

JIAN-GUO LIU (*Department of Mathematics and Department of Physics, Duke University, Durham NC 27708, USA*),

AND

YULONG LU (*Department of Mathematics, Duke University, Durham NC 27708, USA*)

Abstract. We consider a sequence of identical independently distributed random samples from an absolutely continuous probability measure in one dimension with unbounded density. We establish a new rate of convergence of the $\infty-$Wasserstein distance between the empirical measure of the samples and the true distribution, which extends the previous convergence result by Trillos and Slepčev to the case that the true distribution has an unbounded density.

1. Introduction. Consider a sequence of identical independently distributed (i.i.d.) random variables $\{X_i\}, i = 1, \cdots, n$, sampled from a given probability measure $\nu \in \mathscr{P}(\mathbb{R}^d)$ with probability density function ρ. Here $\mathscr{P}(\mathbb{R}^d)$ denotes the space of all probability measures on $\mathbb{R}^d$. We define the empirical measure ν_n associated to the samples $\{X_i\}$ by

$$\nu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}.$$

The well-known Glivenko–Cantelli theorem [20] states that ν_n converges weakly to ν as $n \to \infty$. In recent years, there has been growing interest in quantifying the rate of convergence of ν_n to ν with respect to Wasserstein distances. Recall that the p-Wasserstein distance between two probability measures $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$ is defined as

$$W_p(\mu,\nu) := \Big(\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} |x-y|^p \Pi(dx,dy)\Big)^{1/p}, \quad 1 \le p < \infty$$

and

$$W_\infty(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \operatorname{esssup}_\pi |x-y|,$$

Received .

E-mail address: lan15@mails.tsinghua.edu.cn
E-mail address: jliu@phy.duke.edu
E-mail address: yulonglu@math.duke.edu

where $\Pi(\mu,\nu)$ is the set of all probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with two marginals μ and ν.

The purpose of this paper is to prove the rate of convergence of ν_n to ν w.r.t. ∞-Wasserstein distance W_∞ when the density function ρ of ν is unbounded. For simplicity, we will focus on the one dimensional case, but the arguments of the proof are expected to be generalized to high dimensions.

1.1. *Motivation and Related Work.* Estimating the distance between the empirical measure of a sequence of i.i.d. random variables and its true distribution is a highly important problem in probability and statistics. For example, in statistics, it is usually impossible to access to the true distribution, e.g. the posterior distribution in a Bayesian procedure. So in order to extract useful information from the true distribution, a common approach is to generate i.i.d samples from the true distribution via various sampling algorithms (Markov chain Monte Carlo for instance), from which one can approximately compute many statistical quantities of interest, such as the mean or variance by their empirical counterparts. Hence understanding the statistical error in estimating the statistics requires a quantification of the distance between the empirical measure and the true distribution.

The Wasserstein distance is a natural choice for measuring the closeness of two probability measures in the problem of consideration since it allows the probability measures to be singular to each other, which typically allows including Dirac masses or the empirical measures. This is prohibited if total variation distance or Hellinger distance[14] are used. We are particularly interested in the ∞-Wasserstein distance for several reasons. First, the ∞-Wasserstein distance $W_\infty(\mu,\nu)$ reduces to the so-called min-max matching distance [1, 2, 15] when both μ and ν are discrete measures with the same number of Diracs. Such min-max matching distance plays an important role in the analysis of shape matching problems in computer vision; see [9] and the references therein. Moreover, the ∞-Wasserstein distance is also useful in understanding the asymptotic performance of spectral clustering [18, 19]. In fact, in [18], the authors studied the consistency of spectral clustering algorithms in the large graph limit. By formulating the clustering procedure in a variational framework, they characterized the convergence of eigenvalues, eigenvectors of a weighted graph Laplacian, and that of spectral clustering to their underlying continuum limits using Γ-convergence. One crucial ingredient needed in their proof is exactly a convergence rate estimate on the ∞-Wasserstein distance between the empirical measures and the true distribution which was established in [19]. However, they made a strong assumption that the density function of the true distribution is strictly bounded from above and below. We aim to extend the result in [19] to the case where the true distribution has an unbounded density in one dimensional space.

Let us briefly review some important previous works on the rate of convergence of $W_p(\nu_n,\nu)$ with $p \geq 1$. For $p=1$, it was shown by Dudley in [11] that when $d \geq 2$,

$$C_2 \cdot n^{-\frac{1}{d}} \leq \mathbb{E}(W_1(\nu,\nu_n)) \leq C_1 \cdot n^{-\frac{1}{d}}.$$

Based on Sanov's theorem, Bolley, Guillin and Villani [6] proved a concentration estimate on $W_p(\nu_n,\nu)$ for $1 \leq p \leq 2$ in any dimension

$$\mathbb{P}\big(W_p(\nu_n,\nu) \geq t\big) \leq C \cdot e^{-Knt^2}.$$

Boissard [4] extended this result to more general spaces rather than $\mathbb{R}^d$ when $p = 1$ and applied it to the occupation measure of a Markov chain. In [5], Boissard and Gouic gave the rate of convergence for $\mathbb{E}(W_p(\nu_n,\nu)^p)$ when $1 \le p < \infty$. Fournier and Guillin [13] presented a better result than [6, 4] for non-asymptotic moment estimates and concentration estimates. They showed that if ν has finite q-th moment and $p < \frac{d}{2}$, then

$$\mathbb{E}(W_p^p(\nu,\nu_n)) \le C_{q,p} \cdot n^{-\frac{p}{d}},$$

and

$$\mathbb{P}\big(W_p(\nu_n,\nu) \ge t\big) \le C \cdot \exp(-Knt^{\frac{d}{p}}).$$

(We only list the case $p < \frac{d}{2}$ here. For other cases, one can refer to Theorem 1 and 2 in [13].) Weed and Bach gave a new definition of the upper Wasserstein dimension $d^*(\nu)$ for measure ν. They proved that for $1 \le p < \infty$ and $s < d^*(\nu)$,

$$\mathbb{E}(W_p(\nu,\nu_n)) \le C \cdot n^{-\frac{1}{s}}.$$

As for $W_\infty(\nu,\nu_n)$, its rate of convergence is less studied than that of $W_p(\nu,\nu_n)$ with $p < \infty$. As far as we know, most results on $W_\infty(\nu,\nu_n)$ are obtained when ν and ν_n are both discrete measures. As mentioned above, the $\infty-$ Wasserstein distance between two discrete measures is closely linked to the min-max matching problem. Many results have been obtained for the latter when ν is a uniform distribution. Let $S = [0,1]^d$. Define a regularly spaced array of n grid points on S (with $n = k^d$ for some $k \in \mathbb{N}$) by Y_i and the i.i.d. random samples with uniform distribution on S by X_i. Leighton and Shor [15], and Yukich and Shor [16] showed that as $n \to \infty$, it holds with high probability that

$$W_\infty(\nu,\nu_n) = \min_{\pi}\max_{i} |X_{\pi_i} - Y_i| \sim \begin{cases} O\left(\dfrac{(\log n)^{\frac{3}{4}}}{n^{\frac{1}{2}}}\right), & d = 2, \\ O\left(\left(\dfrac{\log n}{n}\right)^{\frac{1}{d}}\right), & d \ge 3, \end{cases}$$

where π is a permutation of $\{1,2,\cdots,n\}$. When ν has a Lebesgue density which is bounded from above and away from zero, Trillos and Slepčev [19] proved that the above estimate still holds. Davis and Sethuraman [10] showed that in 1-D, the following estimate holds

$$W_\infty(\nu,\nu_n) \le C \cdot \left(\frac{\log(\log(n))}{n}\right)^{\frac{1}{2}},$$

provided that the density function is bounded above and below by positive constants and satisfies additional Lipschitz condition. We will prove a similar result in Theorem 1.1 without the upper bound assumption on the density.

1.2. *Main Results.* The purpose of this paper is to improve the results of [19] in 1-D by removing the boundedness constraint on $\rho(x)$. Our first result is a rate of convergence result in the case where the density function $\rho(x)$ is bounded from below, but not from above.

w_1 THEOREM 1.1. Let $D=(0,1)\subseteq\mathbb{R}$ and ν be a probability measure in D with a density function $\rho : D\to(0,\infty)$. Assume that there exists a constant $\lambda\in(0,1)$ such that

$$\rho(x)\geq\lambda,\ \forall\, x\in D.$$

Let $X_1,\cdots,X_n,\cdots$ be i.i.d. random variables sampled from ν and let ν_n be the corresponding empirical measure. Then for any $t>0$,

$$\mathbb{P}\left(W_\infty(\nu,\nu_n)\geq\frac{t}{\lambda}\right)\leq 2\exp(-2nt^2).$$

In particular, except on a set with probability $2n^{-2}$,

$$W_\infty(\nu,\nu_n)\leq\frac{1}{\lambda}\left(\frac{\log n}{n}\right)^{\frac{1}{2}}. \tag{1}$$

result_w_1

REMARK 1.1. Note that the right hand side of (1) will blow up if $\lambda\to 0$. That's why we assume that $\rho(x)$ has a uniform positive lower bound in Theorem 1.1. Moreover, the exponent one half is sharp owing to the central limit theorem.

We proceed to discussing the case when the density function is not strictly bounded away from zero. We first comment that if the density function of ν is zero in a connected region, then by definition the ∞-Wasserstein distance between ν_n and ν can not go to zero as n goes to infinity. In fact, consider the probability measure ν_0 with the density function

$$\rho_0(x)=\begin{cases}\dfrac{3}{2},\, x\in\left(0,\dfrac{1}{3}\right)\bigcup\left(\dfrac{2}{3},1\right),\\ 0,\, x\in\left[\dfrac{1}{3},\dfrac{2}{3}\right].\end{cases}$$

Let $\nu_{n,0}$ be the empirical measure of ν_0. Since $\nu_{n,0}$ depends on a sequence of random variables, there is no guarantee that $\nu_{n,0}((0,\frac{1}{3}))=\nu_0((0,\frac{1}{3}))$. Assume that $\nu_{n,0}((0,\frac{1}{3}))=\nu_0((0,\frac{1}{3}))+\delta_n$, where $1\gg\delta_n>0$ is a small parameter. Since $W_\infty(\nu_{n,0},\nu_n)$ is also the maximal distance that an optimal transportation map from $\nu_{n,0}$ to ν_0 moves the mass by (which will be mentioned later in Lemma 2.2), it follows that

$$W_\infty(\nu_{n,0},\nu_0)\geq\operatorname{diam}\left(\left(\frac{1}{3},\frac{2}{3}\right)\right)=\frac{1}{3}>0.$$

Therefore, for the validity of Theorem 1.2, we assume that $\rho(x)$ satisfies the following conditions.

(A-1) There are only N points $x_1,\cdots,x_N$ in D satisfying $\rho(x_i)=0$.

(A-2) For each zero point x_i,

$$m_i|x_i-x|^{k_i}\leq\rho(x),\ \text{for}\ \forall\, x\in B_i, \tag{2}$$

control_rho

where $B_i=(x_i-\Delta_i,x_i+\Delta_i)$ is a small neighborhood of x_i and Δ_i,k_i,m_i are positive numbers.

w_2 THEOREM 1.2. Let $D = (0,1) \subseteq \mathbb{R}$ and ν be a probability measure in D with a density function $\rho : D \to (0,\infty)$. Assume that ρ satisfies Assumptions (A-1) and (A-2). Assume further that there exists a constant $\Lambda > 0$ such that for all $x \in D$,

$$\rho(x) \leq \Lambda.$$

Then there exists a positive constant $C = C(k_i, m_i)$ such that except on a set with probability $O\left(\frac{1}{\log n}\right)$,

$$W_\infty(\nu, \nu_n) \leq C \cdot \max_i \left(\frac{\log n}{n}\right)^{\frac{1}{2(k_i+1)}}.$$

We would like to sketch the proof of the theorems above. To prove Theorem 1.1, we use the fact that in one dimension, W_∞ distance between two measures can be written as the L^∞ norm of the difference of their quantile functions. Moreover, thanks to the $\frac{1}{\lambda}$-Lipschitz continuity of the quantile function of ν, which follows by the assumption that $\rho \geq \lambda$, the L^∞ norm of the difference of the quantile functions can be bounded from above by the difference between the cumulative distribution function of the true distribution ν and that of the empirical distribution ν_n. Finally, the latter can be bounded by using the Dvoretzky-Kiefer-Wolfowitz inequality [12].

For the proof of Theorem 1.2, we first divide the domain D into a family of sub-domains according to the value of $\rho(x)$. Then, we use the following scaling equality in each sub-domain

$$W_\infty(\nu, \nu_n) = W_\infty(\theta\nu, \theta\nu_n),$$

with an appropriate scaling parameter θ such that after rescaling, the Lebesgue density of the rescaled measure $\theta\nu$ is bounded from above and below. With the density being bounded, we can estimate the ∞-Wasserstein distance by using the same method in [19]. However, the mass of ν and ν_n may not be equal in each sub-domain. To resolve this issue, we introduce a new measure $\widetilde{\nu}$ such that $\widetilde{\nu}$ has the same mass as ν_n in each sub-domain. Since the distance between $\widetilde{\nu}$ and ν_n can be bounded by an argument similar to Theorem 1.1 in [19], it suffices to estimate the distance between ν and $\widetilde{\nu}$.

The following corollary is a direct consequence of Theorem 1.1 and Theorem 1.2.

w_3 COROLLARY 1.1. Let $D = (0,1) \subseteq \mathbb{R}$ and ν be a probability measure in D with density $\rho : D \to (0,\infty)$. Assume that $\rho(x)$ satisfies (A-1), (A-2). Let $X_1, \cdots, X_n, \cdots$ be i.i.d. random variables sampled from ν. Then there exists a positive constant $C = C(\delta, M, k_i, m_i)$ such that except on a set with probability $O\left(\frac{1}{\log n}\right)$,

$$W_\infty(\nu, \nu_n) \leq C \cdot \max_i \left(\frac{\log n}{n}\right)^{\frac{1}{2(k_i+1)}}. \tag{3}$$

eq:rate

1.3. *Discussion.* As we mentioned earlier, quantifying the rate of convergence of ν_n to ν with respect to $\infty-$Wasserstein distance is very useful for understanding the consistency of spectral clustering[18]. Our new convergence rate estimates will reshape the convergence of spectral clustering in the case where the density of true distribution is unbounded, as we discuss in what follows.

Let $V = \{x_1, \cdots, x_n\}$ be a set of data points in $\mathbb{R}^d$ sampled from a probability measure ν. For each pair of points x_i and x_j, we construct a weight $W_{i,j}^{\varepsilon_n}$ between them to characterize their similarities. In general, the weight has the form of

$$W_{i,j}^{\varepsilon_n} = \eta_{\varepsilon_n}(x_i - x_j),$$

where $\eta_{\varepsilon_n}(z) = \frac{1}{\varepsilon_n^d}\eta(\frac{z}{\varepsilon_n})$ and η is an appropriate kernel function(for example, Gaussian kernel). The weight matrix $W^{\varepsilon_n} \in \mathbb{R}^{n\times n}$ is then defined by $W_{i,j}^{\varepsilon_n}$. Let $D^{\varepsilon_n} \in \mathbb{R}^{n\times n}$ be a diagonal matrix with $D_{ii}^{\varepsilon_n} = \sum_j W_{i,j}^{\varepsilon_n}$. Then the discrete Dirichlet energy and the relevant continuum Dirichlet energy are defined by

$$G_{n,\varepsilon_n}(u^n) = \frac{1}{\varepsilon_n^2 n^2}\sum_{i,j} W_{i,j}^{\varepsilon_n}\left(u^n(x_i) - u^n(x_j)\right)^2, \ u^n \in L^2(\nu_n),$$

and

$$G(u) = \int_D |\nabla u|^2 \rho^2(x)dx, \ u \in L^2(\nu),$$

where $\rho(x)$ is the density function of the underlying measure ν. The unnormalized graph Laplacian L_{n,ε_n} is defined by

$$L_{n,\varepsilon_n} = D^{\varepsilon_n} - W^{\varepsilon_n}.$$

The aim of spectral clustering is to partition the data points $x_1, \cdots, x_n$ into k meaningful groups. To do this, the spectrum of unnormalized graph Laplacian L_{n,ε_n} is used to embed the data points into a low dimensional space. Then we can apply some clustering algorithms like k-means to these points. For more details about spectral clustering, one can see [22].

In [18], the authors proved that when the density function $\rho(x)$ of ν is bounded from above and below, the spectrum of unnormalized graph Laplacian L_{n,ε_n} converges to the spectrum of the corresponding continuum operator L, which implies the consistency of spectral clustering. They also gave a lower bound of the convergence rate at which the connectivity radius $\varepsilon_n \to 0$ as $n \to \infty$. With our theorems, the results in [18] can be generalized to the case when $\rho(x)$ is unbounded. In particular, the kernel width ε_n should be chosen to be slightly bigger than the right side of (3), which is different from [18].

The proof will not be included in this paper since it is similar to the proof in [18]. We sketch the outline of the proof as follows:

First, we prove the Γ-convergence of Dirichlet energy G_{n,ε_n} to G. Note that $G_{n,\varepsilon_n}(u^n)$ and $G(u)$ are defined in different function spaces. In order to show the Γ-convergence, we need to construct an approximate function to $u^n \in L^2(\nu_n)$ in $L^2(\nu)$ by

$$\widetilde{u} = u_n \circ T_n,$$

where T_n is the transportation plan between ν and ν_n. Our theorems are used in this step to establish the probabilistic estimates and the constraint on ε_n. By corollary 1.1 and lemma 2.2

below, we can give an upper bound on $\|T_n - I\|$

$$\limsup_{n\to\infty}\min_i \frac{n^{\frac{1}{2(k_i+1)}}\|T_n - I\|}{(\log n)^{\frac{1}{2(k_i+1)}}} \le C.$$

Then by requiring that $\varepsilon_n \gg \max_i \left(\frac{\log n}{n}\right)^{\frac{1}{2(k_i+1)}}$ we can give upper and lower bounds on $G_{n,\varepsilon_n}(u^n)$ in terms of $G(\widetilde{u})$. Hence the Γ- convergence of Dirichlet energy G_{n,ε_n} to G can be proved.

Next, by min-max theorem, we know that the eigenvalues of L_{n,ε_n}(or L) can be written as the minimizers of G_{n,ε_n}(or G). Therefore, the convergence of spectrum is equivalent to the convergence of the minimizers of G_{n,ε_n} , which can be proved by the Γ- convergence and compactness properties of G_{n,ε_n}. Finally, with the convergence of spectrum, we can prove the consistency of spectral clustering.

The paper is organized as follows: In section 2, we introduce some preliminaries and notations. In section 3.1 and section 3.2, we prove Theorem 1.1 and Theorem 1.2 respectively. Finally, the proof of Corollary 1.1 is presented in section 3.3.

2. Preliminaries and notations.

2.1. *Notations.* Let $D = (0,1) \subset \mathbb{R}$ and $\mathscr{P}(D)$ be the set of all probability measures on D. Given a probability measure $\mu \in \mathscr{P}(D)$ and a Borel-measurable map T, we define the pushforward ν of measure μ under the map T by setting

$$\nu(A) = T_\sharp\mu(A) = \mu(T^{-1}(A))$$

for any measurable set $A \subset D$. We call T the transportation map between μ and ν.

The $\infty-$Wasserstein distance $W_\infty(\mu,\nu)$ is defined by

$$W_\infty(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \operatorname{esssup}_\pi |x-y|,$$

where $\Pi(\mu,\nu)$ is the set of all couplings between μ and ν, i.e.

$$\Pi(\mu,\nu) = \Big\{\pi \in \mathscr{P}((0,1)^2) | \pi(A\times(0,1)) = \mu(A),$$
$$\pi((0,1)\times B) = \nu(B), \text{ for all Borel sets } A, B \subset (0,1)\Big\}.$$

REMARK 2.1. Note that the definition of $W_\infty(\mu,\nu)$ can be generalized to the case where μ and ν have the same mass on D. Therefore, in the sequential, we still write $W_\infty(\mu,\nu)$ when $\mu(D) = \nu(D)$ even though μ and ν are not necessarily probability measures.

It was proved in [7] that if μ is absolutely continuous with respect to the Lebesgue measure, then for any optimal transport plan π of $W_\infty(\mu,\nu)$, there exists a transportation map $T : D \to D$ such that $T_\sharp\mu = \nu$ and $\pi = (I\times T)_\sharp\mu$. In particular, the optimal transportation plan of $W_\infty(\nu,\nu_n)$, with ν_n being the empirical measure of the absolutely continuous probability measure ν is unique.

2.2. *Useful lemmas.* The following lemma collects some properties on W_∞ to be used in subsequent sections. The proof is trivial and thus is omitted.

infiniteW

LEMMA 2.1. Given measures μ_1, μ_2, μ_3 defined on D with $\mu_1(D) = \mu_2(D) = \mu_3(D)$, then the followings hold:

(1) Triangle inequality: $W_\infty(\mu_1, \mu_3) \le W_\infty(\mu_1, \mu_2) + W_\infty(\mu_2, \mu_3)$.
(2) Scaling equality: $W_\infty(\mu_1, \mu_2) = W_\infty(\alpha\mu_1, \alpha\mu_2)$, for $\forall \alpha > 0$.
(3) $W_\infty(\mu_1, \mu_2) \le \mathrm{diam}(D)$.
(4) If $D = \bigsqcup_j D_j$ then

$$W_\infty(\mu_1, \mu_2) \le \max_j W_\infty(\mu_1|_{D_j}, \mu_2|_{D_j}).$$

The following two lemmas gives two different characterizations of $W_\infty(\mu, \nu)$.

Anoth_def

LEMMA 2.2 ([7]). Let μ, ν be two Borel measures with μ absolutely continuous with respect to the Lebesgue measure and $\mu(D) = \nu(D)$. Then there exists an optimal transportation map $T : D \to D$ such that $T_\sharp \mu = \nu$ and

$$W_\infty(\mu, \nu) = \|I - T\|_{L^\infty(D)}.$$

Furthermore, if $\nu = \Sigma_{i=1}^k a_i \delta_{y_i}$ with $y_i \in D$ and positive numbers $a_i, i = 1, \cdots, k$, then there exists a unique transportation map $T^\star : D \to D$ such that

$$W_\infty(\mu, \nu) = \|I - T^\star\|_{L^\infty(D)}.$$

lemma_quan

LEMMA 2.3 ([21, Remark 2.19]). Let μ, ν be two probability measures on $\mathbb{R}$. Denote the cumulative distribution functions of μ and ν by $F(x)$ and $G(x)$ respectively. Then we have the following equality that

$$W_\infty(\mu, \nu) = \left\|F^{-1} - G^{-1}\right\|_{L^\infty}.$$

inf_den

LEMMA 2.4 ([19, Lemma 2.2]). Let ν_1 and ν_2 be two probability measures defined on D with density functions $\rho_1(x)$ and $\rho_2(x)$ respectively. Assume that there exists a positive constant $\lambda > 0$ such that

$$\rho_i(x) \ge \lambda > 0, \; i = 1, 2.$$

Then there exists $C > 0$ such that

$$W_\infty(\nu_1, \nu_2) \le \frac{C}{\lambda} \cdot \mathrm{diam}(D) \|\rho_1(x) - \rho_2(x)\|_{L^\infty(D)}.$$

The following three probability inequalities on binomial random variables and the Dvoretzky-Kiefer-Wolfowitz inequality will be used in the proofs of main results.

LEMMA 2.5. Let $S_n \sim \operatorname{Bin}(n,p)$ be the independent binomial random variables. For $t > 0$, Chebychev's inequality[17] states that

$$\mathbb{P}\left(\frac{|S_n - n\cdot p|}{\sqrt{np(1-p)}} \geq t\right) \leq \frac{1}{t^2}.$$

The Chernoff's inequality [8] states that

$$\mathbb{P}\Big(\big|\frac{S_n}{n} - p\big| \geq t\Big) \leq 2\exp(-2nt^2).$$

Bernstein's inequality [3] states that

$$\mathbb{P}\Big(\big|\frac{S_n}{n} - p\big| \geq t\Big) \leq 2\exp\Big(-\frac{\frac{1}{2}n^2t^2}{np(1-p)+\frac{1}{3}nt}\Big).$$

LEMMA 2.6 (Dvoretzky-Kiefer-Wolfowitz inequality [12]). Let $\{X_i\}_{i=1}^n$ be the i.i.d. random variables sampled from a probability measure ν. Let $F(x)$ be the cumulative distribution function of ν and $F_n(x)$ be the cumulative distribution function of ν_n. Then for $\forall t > 0$,

$$\mathbb{P}\left(\sup_x |F_n(x) - F(x)| \geq t\right) \leq 2\exp(-2nt^2).$$

3. Convergence of empirical measure.

3.1. *Proof of Theorem 1.1.*

Proof. Denote the cumulative distribution function of ν_n by $F_n(x)$ and that of ν by $F(x)$. Thanks to the Dvoretzky-Kiefer-Wolfowitz inequality [12],

$$\mathbb{P}\left(\sup_x |F_n(x) - F(x)| \geq t\right) \leq 2\exp(-2nt^2).$$

From this, we claim that

$$\mathbb{P}\left(\sup_y \left|F_n^{-1}(y) - F^{-1}(y)\right| \geq \frac{t}{\lambda}\right) \leq \mathbb{P}\left(\sup_x |F_n(x) - F(x)| \geq t\right) \tag{4}$$

claim_1

which implies that

$$\mathbb{P}\left(\sup_y \left|F_n^{-1}(y) - F^{-1}(y)\right| \geq \frac{t}{\lambda}\right) \leq 2\exp(-2nt^2).$$

To prove (4), it suffices to show that $\sup_x |F_n(x) - F(x)| \leq t$ implies $\sup_y \left|F_n^{-1}(y) - F^{-1}(y)\right| \leq \frac{t}{\lambda}$. To this end, fix $y \in [0,1]$. Let $x_1 = F_n^{-1}(y)$ and $x_2 = F^{-1}(y)$. Then from the fact that the density function $\rho(x)$ has a lower bound λ we know that

$$\frac{|F(x_1) - F(x_2)|}{|x_1 - x_2|} \geq \lambda.$$

It follows that

$$\lambda|x_1 - x_2| \leq |F(x_1) - F(x_2)| = |F(x_1) - F_n(x_1)| \leq t,$$

where the last inequality is obtained from $\sup_x |F_n(x) - F(x)| \leq t$. Therefore, for any y,

$$\left|F_n^{-1}(y) - F^{-1}(y)\right| \leq \frac{t}{\lambda}.$$

which completes the proof of (4). It follows from (4) and Lemma 2.3 that

$$\mathbb{P}\left(W_\infty(\nu,\nu_n)\geq\frac{t}{\lambda}\right)\leq 2\exp(-2nt^2).$$

By taking $t=\left(\frac{\log n}{n}\right)^{\frac{1}{2}}$ we get that except on a set with probability $2n^{-2}$,

$$W_\infty(\nu,\nu_n)\leq\frac{1}{\lambda}\left(\frac{\log n}{n}\right)^{\frac{1}{2}}.$$

□

3.2. *Proof of Theorem 1.2.*

1lamma LEMMA 3.1. If $a>b>0$, then $a^k-b^k\geq(a-b)^k$.

Proof. By induction, we only need to prove that $a^k-b^k\geq(a-b)^k$ implies $a^{k+1}-b^{k+1}\geq(a-b)^{k+1}$. From $a>b>0$ we know $2b^{k+1}\leq ab^k+ba^k$. Therefore,

$$(a-b)^{k+1}=(a-b)(a-b)^k\leq(a-b)(a^k-b^k)=a^{k+1}+b^{k+1}-ab^k-ba^k\leq a^{k+1}-b^{k+1}. \quad (5)$$

□

In Theorem 1.2, we give the rate of convergence of $W_\infty(\nu_n,\nu)$ when the density function $\rho(x)$ is not strictly bounded away from zero. The proof is a refinement of the proof of [19, Theorem 1.1], which deals with the case where $\rho(x)$ is bounded. We sketch the rough idea of our proof in the followings before we give the details.

To prove the theorem, we would like to use Lemma 2.1-(4) to reduce the estimate of $W_\infty(\nu,\nu_n)$ to that of $W_\infty(\nu|_{B_i},\nu_n|_{B_i})$, where B_i is a small neighborhood of the zero point x_i . For doing so, we need to modify the measure ν locally (denote the new measure to be $\tilde{\nu}$ after modification) so that $\tilde{\nu}$ has the same mass as ν_n on B_i. Then, we divide B_i into a family of sub-domains $\{A_j\}_{j\in\mathbb{N}}$ according to the value of $\rho(x)$ so that ρ is bounded from above and below on A_j. Thus we can adapt similar arguments from [19] to obtain bounds on $W_\infty(\widetilde{\nu}|_{A_j},\nu_n|_{A_j})$. However, ν_n may not have the same mass as $\widetilde{\nu}$ on each A_j. So, in order to remove this mass discrepancy, we introduce another new measure $\overline{\nu}$ such that $\overline{\nu}(A_j)=\nu_n(A_j)$. At last, thanks to Lemma 2.1, we can establish an upper bound on $W_\infty(\overline{\nu}|_{B_i},\nu_n|_{B_i})$ with the estimates of $W_\infty(\overline{\nu}|_{A_j},\nu_n|_{A_j})$.

Proof. Let $B_{N+1}=(0,1)\backslash\cup_1^N B_i$. Then $\{B_i\}_{i=1}^{N+1}$ is a partition of D. Let $\varepsilon_i=\frac{\nu_n(B_i)}{\nu(B_i)}-1$ for $i=1,\cdots,N+1$ and $\widetilde{\nu}$ be a probability measure defined on D

$$d\widetilde{\nu}=\left(\sum_{i=1}^{N+1}(1+\varepsilon_i)\rho(x)\mathbb{1}_{B_i}\right)dx. \quad (6)$$

nu:widetilde

Then it's clear that

$$\widetilde{\nu}(B_i)=(1+\varepsilon_i)\nu(B_i)=\nu_n(B_i).$$

Combining this with Lemma 2.1, we obtain that

$$W_\infty(\nu,\nu_n)\leq W_\infty(\nu,\widetilde{\nu})+W_\infty(\widetilde{\nu},\nu_n)\leq W_\infty(\nu,\widetilde{\nu})+\max_{i=1,\cdots,N+1}W_\infty(\widetilde{\nu}|_{B_i},\nu_n|_{B_i}).$$

Choose $\beta>2$. To estimate $W_\infty(\widetilde{\nu}|_{B_i},\nu_n|_{B_i})(i=1,\cdots,N)$, we divide B_i into a family of sub-domains $\{A_j\}_{j\in\mathbb{N}}$ and use scaling property to bound W_∞ distance on each sub-domain A_j.

Define $\{A_j\}_{j\in\mathbb{N}}$ by $A_0 = \{x : 1 < \rho(x) \leq \Lambda\}\bigcap B_i$, $A_j = \left\{x : \frac{1}{(j+1)^\beta} < \rho(x) \leq \frac{1}{j^\beta}\right\}\bigcap B_i$ (If A_j is empty, just neglect it). Then,

$$B_i = \bigsqcup_j A_j.$$

Let $\delta_j = \frac{\nu_n(A_j)}{\nu(A_j)} - 1$ and define a measure $\overline{\nu}$ on B_i by

$$d\overline{\nu} = \sum_j \mathbb{1}_{A_j}(1+\delta_j)\rho(x)dx.$$

Then it's easy to see that

$$\overline{\nu}(A_j) = (1+\delta_j)\nu(A_j) = \nu_n(A_j).$$

Again, with this and Lemma 2.1, we can bound $W_\infty(\widetilde{\nu}|_{B_i}, \nu_n|_{B_i})(i = 1, \cdots, N)$ as follows

$$\begin{aligned} W_\infty(\widetilde{\nu}|_{B_i}, \nu_n|_{B_i}) &\leq W_\infty(\widetilde{\nu}|_{B_i}, \overline{\nu}|_{B_i}) + W_\infty(\overline{\nu}|_{B_i}, \nu_n|_{B_i}) \\ &\leq W_\infty(\widetilde{\nu}|_{B_i}, \overline{\nu}|_{B_i}) + \sup_j W_\infty(\overline{\nu}|_{A_j}, \nu_n|_{A_j}). \end{aligned}$$

Therefore, to estimate $W_\infty(\nu, \nu_n)$, it suffices to estimate $W_\infty(\nu, \widetilde{\nu})$, $W_\infty(\widetilde{\nu}|_{B_i}, \overline{\nu}|_{B_i})$, $W_\infty\left(\overline{\nu}|_{A_j}, \nu_n|_{A_j}\right)$, and $W_\infty\left(\widetilde{\nu}|_{B_{N+1}}, \nu_n|_{B_{N+1}}\right)$ respectively.

Step 1: We first estimate $W_\infty\left(\widetilde{\nu}|_{B_{N+1}}, \nu_n|_{B_{N+1}}\right)$. It's easy to deduce, via Lemma 2.1, that

$$\begin{aligned} W_\infty\left(\widetilde{\nu}|_{B_{N+1}}, \nu_n|_{B_{N+1}}\right) &= W_\infty\left(\frac{1}{\widetilde{\nu}(B_{N+1})}\widetilde{\nu}|_{B_{N+1}}, \frac{1}{\widetilde{\nu}(B_{N+1})}\nu_n|_{B_{N+1}}\right) \\ &= W_\infty\left(\frac{1}{\nu(B_{N+1})}\nu|_{B_{N+1}}, \frac{1}{\sum\delta_{X_i}(B_{N+1})}\sum_{X_i\in B_{N+1}}\delta_{X_i}|_{B_{N+1}}\right). \end{aligned}$$

To ease the notations, we write $\nu_{N+1} = \frac{1}{\nu(B_{N+1})}\nu|_{B_{N+1}}$ and $\nu_{n,N+1} = \frac{1}{\sum\delta_{X_i}(B_{N+1})}\sum_{X_i\in B_{N+1}}\delta_{X_i}|_{B_{N+1}}$.

Clearly, ν_{N+1} is the restriction of ν to B_{N+1} and $\nu_{n,N+1}$ is the empirical measure of ν_{N+1}. Furthermore, we note that $\rho(x)$ is bounded from below in B_{N+1} due to the fact that $B_i(i = 1, \cdots, N)$ is a small neighborhood of zero point x_i and $B_{N+1} = D\setminus\bigcup_1^N B_i$. Therefore, we can use Theorem 1.1 to give an estimate on $W_\infty\left(\widetilde{\nu}|_{B_{N+1}}, \nu_n|_{B_{N+1}}\right)$.(We remark that Theorem 1.1 holds true for any domain $(a,b)\subset\mathbb{R}$ by replacing $D = (0,1)$ with $D = (a,b)$ in the proof.)

Let $\lambda_{N+1} := \min_{x\in B_{N+1}}\rho(x)$. Then we have $0 < \frac{\lambda_{N+1}}{\nu(B_{N+1})} \leq \frac{1}{\nu(B_{N+1})}\rho(x)|_{B_{N+1}}$. It follows from Theorem 1.1 that there exists a constant $C = \frac{\nu(B_{N+1})}{\lambda_{N+1}}$ such that

$$W_\infty\left(\widetilde{\nu}|_{B_{N+1}}, \nu_n|_{B_{N+1}}\right) \leq C\left(\frac{\log n}{n}\right)^{\frac{1}{2}}.$$

Step 2: We then estimate $W_\infty\left(\overline{\nu}|_{A_j}, \nu_n|_{A_j}\right)$. To achieve this, set

$$J_0 = \left\lfloor\left(\frac{n}{\log n}\right)^{\frac{k_i}{2\beta(k_i+1)}}\right\rfloor - 1$$

and consider the following two cases: 1) $j < J_0$ and 2) $j \geq J_0$.

We claim that, when $j \geq J_0$, $\operatorname{diam}(A_j) \leq C\cdot\left(\frac{\log n}{n}\right)^{\frac{1}{2(k_i+1)}}$. To show the claim, we first recall the definition that $A_j = \left\{x : \frac{1}{(j+1)^\beta} < \rho(x) \leq \frac{1}{j^\beta}\right\}\bigcap B_i$ and the assumption that $m_i|x_i - x|^{k_i} \leq \rho(x)$

in B_i. To simplify the notations, we denote $m_i|x-x_i|^{k_i}$ by $\rho_1(x)$. Let x_R be a positive constant satisfying $\rho_1(x_R)=\frac{1}{j^\beta}$.

From $\rho_1(x)\le\rho(x)$ we know that $\operatorname{diam}(A_j)\le x_R$. Moreover, when n is large enough,

$$\begin{aligned}
x_R=\frac{1}{m_i^{\frac{1}{k_i}}}\cdot j^{-\frac{\beta}{k_i}}\le C\cdot J_0^{-\frac{\beta}{k_i}} &= C\cdot\left(\left\lfloor\left(\frac{n}{\log n}\right)^{\frac{k_i}{2\beta(k_i+1)}}\right\rfloor-1\right)^{-\frac{\beta}{k_i}}\\
&\le C\cdot\left(\frac{1}{2}\cdot\left(\frac{n}{\log n}\right)^{\frac{k_i}{2\beta(k_i+1)}}\right)^{-\frac{\beta}{k_i}}\\
&\le C\cdot\left(\frac{\log n}{n}\right)^{\frac{1}{2(k_i+1)}},
\end{aligned}$$

where $C=C(k_i,m_i,\beta)$.

Therefore, when $j\ge J_0$, $\operatorname{diam}(A_j)\le C\cdot\left(\frac{\log n}{n}\right)^{\frac{1}{2(k_i+1)}}$. By Lemma 2.1, when $j\ge J_0$,

$$W_\infty(\overline{\nu}|_{A_j},\nu_n|_{A_j})\le\operatorname{diam}(A_j)\le C\cdot\left(\frac{\log n}{n}\right)^{\frac{1}{2(k_i+1)}},$$

where $C=C(k_i,\lambda_i,\beta)$.

We then turn to the case that $j<J_0$. We first use scaling equality $W_\infty(\overline{\nu}|_{A_j},\nu_n|_{A_j})=W_\infty(j^\beta\overline{\nu}|_{A_j},j^\beta\nu_n|_{A_j})$. For simply notations, let

$$\nu_{n,j}=j^\beta\nu_n|_{A_j},\ \overline{\nu}_j=j^\beta\overline{\nu}|_{A_j}.$$

Then the density function of $\overline{\nu}_j$ is defined by

$$\overline{\rho}_j(x)=j^\beta(1+\delta_j)\rho(x).$$

For every $k\in\mathbb{N}$, we partition A_j into 2^k sub-domains. Each of them have a $\overline{\nu}_j-$mass of $\frac{1}{2^k}\overline{\nu}_j(A_j)$. Let $\mathscr{F}_{k,j}$ be the set of these sub-domains. $\mathscr{F}_{0,j}=A_j$. And $\mathscr{F}_{k+1,j}$ is obtained by bisecting each box in $\mathscr{F}_{k,j}$, according to $\overline{\nu}_j$. Thus, for any $Q\in\mathscr{F}_{k,j}$,

$$\overline{\nu}_j(Q)=\frac{1}{2^k}\overline{\nu}_j(A_j),\ \nu(Q)=\frac{1}{2^k}\nu(A_j).$$

We define a series of new measures $\{\mu_{k,j}\}$ by setting $d\mu_{k,j}(x)=\rho_{k,j}(x)dx$ with

$$\rho_{k,j}(x)=\frac{\nu_{n,j}(Q)}{\overline{\nu}_j(Q)}\cdot\overline{\rho}_j(x)=\frac{\nu_n(Q)}{\nu(Q)}j^\beta\rho(x),\ \forall\,x\in Q\in\mathscr{F}_{k,j}.$$

We claim that for $\forall Q\in\mathscr{F}_{k,j}$, $\forall\ k\le k_n=\log_2\left(\frac{n\nu(A_j)}{10\log n}\right)$, there exists a constant C such that the following inequality holds true with probability at least $1-2n^{-1}$

$$W_\infty(\mu_{k,j}|_Q,\mu_{k+1,j}|_Q)\le C\cdot(j+1)^\beta\cdot\left(\frac{\nu(A_j)\log n}{2^kn}\right)^{\frac{1}{2}}. \tag{7}$$

claim_2

Assume that the claim holds. Note that $\operatorname{diam}(Q) = \int_Q dx \leq \int_Q (j+1)^\beta \rho(x)dx = (j+1)^\beta \nu(Q)$. Then for $j = 1, \cdots, J_0 - 1$, we have

$$\begin{aligned}
W_\infty(\nu_n|_{A_j}, \overline{\nu}|_{A_j}) = W_\infty(\nu_{n,j}, \overline{\nu_j}) &\leq \sum_{k=1}^{k_n} W_\infty(\mu_{k-1,j}, \mu_{k,j}) + W_\infty(\mu_{k_n,j}, \nu_{n,j}) \\
&\leq \sum_{k=1}^{k_n} \left(C \cdot (j+1)^\beta \left(\frac{\nu(A_j)\log n}{2^k n} \right)^{\frac{1}{2}} \right) + \max_{Q\in\mathscr{F}_k} \operatorname{diam}(Q) \\
&\leq \left(\sum_{k=1}^{k_n} \left(C \cdot \left(\frac{\nu(A_j)\log n}{2^k n} \right)^{\frac{1}{2}} \right) + \frac{1}{2^{k_n}} \cdot \nu(A_j) \right) \cdot (j+1)^\beta \\
&\leq C \left(\left(\frac{\log n}{n} \right)^{\frac{1}{2}} + C \frac{\log n}{n} \right) (J_0+1)^\beta \\
&= C \cdot \left(\frac{\log n}{n} \right)^{\frac{1}{2}} \cdot \left\lfloor \left(\frac{n}{\log n} \right)^{\frac{k_i}{2\beta(k_i+1)}} \right\rfloor^\beta \\
&\leq C \left(\frac{\log n}{n} \right)^{\frac{1}{2(k_i+1)}}.
\end{aligned}$$

Therefore, for $\forall j \in \mathbb{N}$,

$$W_\infty(\nu_n|_{A_j}, \overline{\nu}|_{A_j}) \leq C \left(\frac{\log n}{n} \right)^{\frac{1}{2(k_i+1)}}$$

where C depends on β, k_i, λ_i.

Now we return to the proof of the claim (7). Actually, from the definition of $\mu_{k,j}$ and $\rho_{k,j}$ it follows that for $\forall\, x \in Q \in \mathscr{F}_{k,j}$,

$$\frac{\nu_n(Q)}{\nu(Q)} \cdot \frac{j^\beta}{(j+1)^\beta} \leq \rho_{k,j}(x) \leq \frac{\nu_n(Q)}{\nu(Q)},$$

and

$$\mu_{k,j}(Q) = \mu_{k+1,j}(Q) = \nu_{n,j}(Q).$$

Therefore, by Lemma 2.1 we know that

$$W_\infty(\mu_{k+1,j}, \mu_{k,j}) \leq \max_{Q\in\mathscr{F}_{k,j}} W_\infty(\mu_{k+1,j}|_Q, \mu_{k,j}|_Q).$$

Let Q_1 be a sub-domain bisected from Q. Then $Q_1 \in \mathscr{F}_{k+1,j}$. According to Lemma 2.4,

$$\begin{aligned}
W_\infty(\mu_{k+1,j}|_Q, \mu_{k,j}|_Q) &= W_\infty\left(\frac{\nu(Q)}{\nu_n(Q)} \mu_{k+1,j}|_Q, \frac{\nu(Q)}{\nu_n(Q)} \mu_{k,j}|_Q \right) \\
&\leq \frac{C}{\rho_{min}} \cdot \operatorname{diam}(Q) \cdot \left| \frac{\nu_n(Q_1)\nu(Q)}{\nu(Q_1)\nu_n(Q)} - 1 \right| \cdot \left\| j^\beta \rho(x) \right\|_{L^\infty(Q)} \\
&= \frac{C}{\rho_{min}} \cdot \operatorname{diam}(Q) \cdot \left| \frac{2\nu_n(Q_1)}{\nu_n(Q)} - 1 \right| \cdot \left\| j^\beta \rho(x) \right\|_{L^\infty(Q)} \\
&\leq \frac{C}{\rho_{min}} \cdot \operatorname{diam}(Q) \cdot \left| \frac{2\nu_n(Q_1)}{\nu_n(Q)} - 1 \right|,
\end{aligned}$$

where

$$\rho_{min} = \frac{j^\beta}{(j+1)^\beta}\min\left\{1, \frac{\nu_n(Q_1)\nu(Q)}{\nu(Q_1)\nu_n(Q)}\right\} = \frac{j^\beta}{(j+1)^\beta}\min\left\{1, \frac{2\nu_n(Q_1)}{\nu_n(Q)}\right\}.$$

To bound $W_\infty(\mu_{k+1,j}|_Q, \mu_{k,j}|_Q)$, it suffices to estimate $\frac{1}{\rho_{min}}$ and $\left|\frac{2\nu_n(Q_1)}{\nu_n(Q)} - 1\right|$ respectively. We first give a probabilistic estimate on $\frac{1}{\rho_{min}}$.

Note that for $\forall Q \in \mathscr{F}_{k,j}$, $\frac{n\nu_{n,j}(Q)}{j^\beta} \sim Bin(n, \nu(Q))$. Thus, we can use Bernstein's inequality and deduce that for $\forall k \le k_n = \log_2\left(\frac{n\nu(A_j)}{10\log n}\right)$,

$$\begin{aligned}\mathbb{P}\Big(|\frac{\nu_{n,j}(Q)}{j^\beta} - \nu(Q)| \ge \frac{1}{2}\nu(Q)\Big) &\le 2\cdot\exp\left(-\frac{\frac{1}{2}n\left(\frac{\nu(Q)}{2}\right)^2}{\nu(Q)\,(1-\nu(Q)) + \frac{1}{3}\cdot\frac{1}{2}\nu(Q)}\right)\\ &\le 2\cdot\exp\left(-\frac{1}{10}\cdot n\cdot\nu(Q)\right)\\ &= 2\cdot\exp\left(-\frac{1}{10}\cdot n\cdot\frac{1}{2^k}\nu(A_j)\right)\\ &\le 2n^{-1}.\end{aligned}$$

That is, with probability at least $1-2n^{-1}$,

$$\left|\frac{1}{j^\beta}\cdot\frac{\nu_{n,j}(Q)}{\nu(Q)} - 1\right| \le \frac{1}{2}.$$

From the definition of $\nu_{k,j}$ we know

$$\frac{3}{2} \ge \frac{\nu_n(Q)}{\nu(Q)} \ge \frac{1}{2},\ \frac{3}{2} \ge \frac{\nu_n(Q_1)}{\nu(Q_1)} \ge \frac{1}{2}. \tag{8}$$

geq12_2

Therefore,

$$\frac{2\nu_n(Q_1)}{\nu_n(Q)} \ge \frac{\nu(Q_1)}{\nu_n(Q)} = \frac{1}{2}\cdot\frac{\nu(Q)}{\nu_n(Q)} \ge \frac{1}{3},$$

and $\frac{1}{\rho_{min}}$ can be bounded with probability at least $1-2n^{-1}$

$$\frac{1}{\rho_{\min}} = \frac{1}{\frac{j^\beta}{(j+1)^\beta}\min\left\{1, \frac{2\nu_n(Q_1)}{\nu_n(Q)}\right\}} \le \frac{3(j+1)^\beta}{j^\beta} \le 3\cdot 3(1+\beta).$$

We then estimate $\left|\frac{2\nu_n(Q_1)}{\nu_n(Q)} - 1\right|$.

Notice that if we set $m = n\cdot\frac{1}{j^\beta}\cdot\nu_{n,j}(Q)$, then

$$m\cdot\frac{\nu_{n,j}(Q_1)}{\nu_{n,j}(Q)} = \sum_1^n \delta_{X_i}(Q_1) \sim Bin\left(m, \frac{\nu(Q_1)}{\nu(Q)}\right) = Bin\left(m, \frac{1}{2}\right).$$

Using Chernoff's inequality we get that

$$\begin{aligned}\mathbb{P}\left(\left|\frac{\nu_{n,j}(Q_1)}{\nu_{n,j}(Q)} - \frac{1}{2}\right| \ge \left(\frac{2^k\log n}{n\nu(A_j)}\right)^{\frac{1}{2}}\right) &\le 2\exp\left(-2\cdot\frac{m2^k\log n}{n\nu(A_j)}\right)\\ &\le 2\exp(-\log n)\\ &= 2n^{-1},\end{aligned}$$

where the last inequality is obtained from (8). Therefore, with probability at least $1-2n^{-1}$,

$$\left|2\frac{\nu_n(Q_1)}{\nu_n(Q)}-1\right| = \left|2\frac{\nu_{n,j}(Q_1)}{\nu_{n,j}(Q)}-1\right| \le 2\left(\frac{2^k\log n}{n\nu(A_j)}\right)^{\frac{1}{2}}.$$

Finally, using the fact that $\operatorname{diam}(Q)=\int_Q dx \le (j+1)^\beta \nu(Q)$, we know that with probability at least $1-2n^{-1}$,

$$\begin{aligned} W_\infty(\mu_{k,j}|_Q, \mu_{k+1,j}|_Q) &\le C\cdot\nu(Q)\left(\frac{2^k\log n}{n\nu(A_j)}\right)^{\frac{1}{2}}\cdot(j+1)^\beta \\ &\le C\cdot(j+1)^\beta\cdot\left(\frac{\nu(A_j)\log n}{2^k n}\right)^{\frac{1}{2}}, \end{aligned}$$

which completes the proof of claim (7).

Step 3: We then estimate $W_\infty(\overline{\nu}|_{B_i}, \widetilde{\nu}|_{B_i})$. We first recall that $B_i=(x_i-\Delta_i, x_i+\Delta_i)$ and

$$d\widetilde{\nu}|_{B_i} = (1+\varepsilon_i)\rho(x)dx\ ,\ d\overline{\nu}|_{B_i} = \sum_j \mathbb{1}_{A_j}(1+\delta_j)\rho(x)dx,$$

where $\varepsilon_i = \frac{\nu_n(B_i)}{\nu(B_i)}-1$ and $\delta_j=\frac{\nu_n(A_j)}{\nu(A_j)}-1$. Let T be the transportation map between $\overline{\nu}|_{B_i}$ and $\widetilde{\nu}|_{B_i}$. Thus for any $x\in B_i$ and $y=Tx$,

$$\int_{x_i-\Delta_i}^{y}\widetilde{\rho}(s)ds = \int_{x_i-\Delta_i}^{x}\overline{\rho}(s)ds.$$

Without loss of generality, we assume $y>x$. Then

$$\int_x^y \widetilde{\rho}(s)ds = \int_{x_i-\Delta_i}^{y}(\overline{\rho}(s)-\widetilde{\rho}(s))ds \le \int_{x_i-\Delta_i}^{y}|\widetilde{\rho}(s)-\overline{\rho}(s)|ds \le \sum_j |\varepsilon_i-\delta_j|\nu(A_j). \tag{9}$$

est_rho_ba

Let $S_n := n\nu_n(A_j)$. Then $S_n=\sum_{i=1}^n \delta_{X_i}(A_j)\sim Bin(n,\nu(A_j))$. According to Chebychev's inequality we know that

$$\mathbb{P}\left(\frac{|S_n-n\cdot\nu(A_j)|}{\sqrt{n\nu(A_j)(1-\nu(A_j))}}\ge\sqrt{\log n}\right)\le(\log n)^{-1},$$

which means that with probability at least $1-(\log n)^{-1}$,

$$\frac{|n\nu_n(A_j)-n\cdot\nu(A_j)|}{\sqrt{n\nu(A_j)(1-\nu(A_j))}}\le\sqrt{\log n}.$$

Then by the definition of δ_j we know that with probability at least $1-(\log n)^{-1}$,

$$|\delta_j\nu(A_j)|\le\Big(\frac{\log n\cdot\nu(A_j)(1-\nu(A_j))}{n}\Big)^{\frac{1}{2}}.$$

With a similar method we derive that with probability at least $1-(\log n)^{-1}$,

$$|\varepsilon_i\nu(B_i)|\le\Big(\frac{\log n\cdot\nu(B_i)(1-\nu(B_i))}{n}\Big)^{\frac{1}{2}}.$$

Note that in A_j, $\rho(s)\le\frac{1}{j^\beta}$, which implies $\nu(A_j)=\int_{A_j}\rho(s)ds\le\int_{A_j}\frac{1}{j^\beta}ds\le\frac{1}{j^\beta}$. Therefore,

$$\sum_j|\delta_j|\nu(A_j)\le\sum_j(\nu(A_j))^{\frac{1}{2}}\left(\frac{\log n}{n}\right)^{\frac{1}{2}}\le\sum_j\frac{1}{j^{\frac{\beta}{2}}}\left(\frac{\log n}{n}\right)^{\frac{1}{2}}\le C\cdot\left(\frac{\log n}{n}\right)^{\frac{1}{2}}.$$

From the fact that $B_i = \bigsqcup_j A_j$ we know

$$\sum_j |\varepsilon_i| \nu(A_j) = |\varepsilon_i| \nu(B_i) \leq \Big(\frac{\log n \cdot \nu(B_i)(1-\nu(B_i))}{n} \Big)^{\frac{1}{2}} \leq C \cdot \left(\frac{\log n}{n} \right)^{\frac{1}{2}}.$$

Therefore from (9) we derive that with probability at least $1-(\log n)^{-1}$,

$$\int_x^y \widetilde{\rho}(s) ds \leq \sum_j |\varepsilon_i - \delta_j| \nu(A_j) \leq \sum_j \big(|\varepsilon_i| + |\delta_j| \big) \nu(A_j) \leq C \cdot \left(\frac{\log n}{n} \right)^{\frac{1}{2}}.$$

Since in B_i, $\widetilde{\rho}(s) \geq (1+\varepsilon_i) m_i |x_i - s|^{k_i}$, it follows that $\int_x^y \widetilde{\rho}(s) ds$ can be bounded from below in the following two cases respectively

$$\int_x^y \widetilde{\rho}(s) ds \geq \begin{cases} (1+\varepsilon_i) m_i \left[(x_i - x)^{k_i+1} + (y - x_i)^{k_i+1} \right], & x_i \in (x,y), \\ (1+\varepsilon_i) m_i \left[(y-x)^{k_i+1} \right], & x_i \notin (x,y). \end{cases}$$

The results are obtained by direct calculations so the proof is omitted here. In both cases, we can derive by lemma 3.1 that

$$|y-x| \leq C \left\{ \left(\frac{\left(\frac{\log n}{n} \right)^{\frac{1}{2}}}{(1+\varepsilon_i) m_i} \right)^{\frac{1}{k_i+1}} \right\}.$$

Therefore,

$$W_\infty(\overline{\nu}|_{B_i}, \widetilde{\nu}|_{B_i}) \leq ||T - I||_{L^\infty(B_i)} \leq \max_{x \in B_i} |y-x| \leq C \left(\frac{\log n}{n} \right)^{\frac{1}{2(k_i+1)}},$$

where C depends on ε_i, m_i and k_i.

Step 4: Finally, for $W_\infty(\nu, \widetilde{\nu})$, we use the same method as step 3 and deduce that

$$W_\infty(\nu, \widetilde{\nu}) \leq C \cdot \left(\frac{\log n}{n} \right)^{\frac{1}{2}},$$

where C depends on ε_i, k_i, m_i.

To sum up, with step 1-4, we know that

$$\begin{aligned} W_\infty(\nu, \nu_n) &\leq W_\infty(\nu, \widetilde{\nu}) + \max \left\{ \max_{i=1,\cdots,N} \left[W_\infty(\widetilde{\nu}|_{B_i}, \overline{\nu}|_{B_i}) + \sup_j W_\infty(\overline{\nu}|_{A_j}, \nu_n|_{A_j}) \right], W_\infty(\widetilde{\nu}|_{B_{N+1}}, \nu_n|_{B_{N+1}}) \right\} \\ &\leq C \cdot \max_i \left(\frac{\log n}{n} \right)^{\frac{1}{2(k_i+1)}}. \end{aligned}$$

where C depends on k_i and m_i. This completes the proof of Theorem 1.2. □

3.3. *Proof of Corollary 1.1.*

Proof. Let $A = \{x : \rho(x) < 1\}$, $B = \{x : \rho(x) \geq 1\}$ and assume that they both are connected sets(otherwise we can divide them into connected sets).

Define a probability measure on D by $d\widetilde{\nu} = ((1+\varepsilon_A)\mathbb{1}_A\rho(x) + (1+\varepsilon_B)\mathbb{1}_B\rho(x))\,dx$, where

$$\varepsilon_A = \frac{\nu_n(A)}{\nu(A)} - 1,\ \varepsilon_B = \frac{\nu_n(B)}{\nu(B)} - 1.$$

Thus, it's easy to see that

$$\widetilde{\nu}(A) = \nu_n(A) \text{ and } \widetilde{\nu}(B) = \nu_n(B). \tag{10}$$

cor_nu

In order to estimate $W_\infty(\nu,\nu_n)$, it suffices to estimate $W_\infty(\nu,\widetilde{\nu})$ and $W_\infty(\widetilde{\nu},\nu_n)$ respectively.

Step 1: We first estimate $W_\infty(\nu_n,\widetilde{\nu})$. Using Lemma 2.1 and (10) we know that

$$\begin{aligned}W_\infty(\widetilde{\nu},\nu_n) &\leq \max\left\{W_\infty\left(\widetilde{\nu}|_A,\nu_n|_A\right), W_\infty\left(\widetilde{\nu}|_B,\nu_n|_B\right)\right\}\\ &= \max\left\{W_\infty\left(\frac{1}{\widetilde{\nu}(A)}\widetilde{\nu}|_A,\frac{1}{\widetilde{\nu}(A)}\nu_n|_A\right), W_\infty\left(\frac{1}{\widetilde{\nu}(B)}\widetilde{\nu}|_B,\frac{1}{\widetilde{\nu}(B)}\nu_n|_B\right)\right\}.\end{aligned}$$

Note that

$$\frac{1}{\widetilde{\nu}(A)}\widetilde{\nu}|_A = \frac{1}{\nu(A)}\nu|_A$$

and

$$\frac{1}{\widetilde{\nu}(A)}\nu_n|_A = \frac{1}{n\widetilde{\nu}(A)}\sum_{i=1}^n \delta_{X_i}|_A = \frac{1}{n\nu_n(A)}\sum_{i=1}^n \delta_{X_i}|_A = \frac{1}{\sum_{i=1}^n \delta_{X_i}(A)}\sum_{X_i\in A}\delta_{X_i}|_A.$$

Therefore, $\frac{1}{\widetilde{\nu}(A)}\nu_n|_A$ is the empirical measure of $\frac{1}{\widetilde{\nu}(A)}\widetilde{\nu}|_A$. By Theorem 1.1 we know that $W_\infty(\widetilde{\nu}|_A,\nu_n|_A) \leq C\cdot\left(\frac{\log n}{n}\right)^{\frac{1}{2}}$.

Similarly, we can deduce that $W_\infty(\widetilde{\nu}|_B,\nu_n|_B) \leq C\cdot\max_i\left(\frac{\log n}{n}\right)^{\frac{1}{2(k_i+1)}}$. Therefore,

$$W_\infty(\widetilde{\nu},\nu_n) \leq C\cdot\max_i\left(\frac{\log n}{n}\right)^{\frac{1}{2(k_i+1)}}.$$

Step 2: We then estimate $W_\infty(\widetilde{\nu},\nu)$.

Let T be the transportation map between $\widetilde{\nu}$ and ν. Then for $\forall x \in D$ and $y = Tx$,

$$\int_0^x \widetilde{\rho}(s)ds = \int_0^y \rho(s)ds.$$

Without loss of generality, we assume $y > x$. Then it follows that

$$\int_x^y \rho(s)ds = \int_0^x \widetilde{\rho}(s) - \rho(s)ds \leq \int_0^x |\widetilde{\rho}(s) - \rho(s)|ds \leq |\varepsilon_A|\nu(A) + |\varepsilon_B|\nu(B).$$

By Chebychev's inequality we know that with probability at least $1-(\log n)^{-1}$,

$$|\varepsilon_A|\nu(A) + |\varepsilon_B|\nu(B) \leq C\left(\frac{\log n}{n}\right)^{\frac{1}{2}}.$$

Thus,

$$\int_{(x,y)\cap A}\rho(s)ds + \int_{(x,y)\cap B}\rho(s)ds \leq C\cdot\left(\frac{\log n}{n}\right)^{\frac{1}{2}}.$$

By the same method in the proof of Theorem 1.2, we can give a lower bound on $\int_{(x,y)\cap A}\rho(s)ds$ and $\int_{(x,y)\cap B}\rho(s)ds$ respectively and conclude that with probability at least $1-(\log n)^{-1}$,

$$W_\infty(\nu,\nu_n) \leq C\cdot\max_i\left(\frac{\log n}{n}\right)^{\frac{1}{2(k_i+1)}}.$$

This completes the proof of Corollary 1.1. □

REMARK 3.1. We showed the rate of convergence of ν_n to ν when the density function $\rho(x)$ is unbounded in one dimension. We expect that similar results also hold to be true in high dimensions. However, the idea of the proof needs to be adapted. In particular, the estimate of $W_\infty(\widetilde{\nu},\nu)$ becomes quite technical in high dimensions, where $\tilde{\nu}$ is an auxiliary measure introduced in (6) for the purpose of removing the mass discrepancy between ν and ν_n in local regions. In fact, in one dimension we estimate $W_\infty(\nu,\widetilde{\nu})$ by using that

$$W_\infty(\nu,\widetilde{\nu}) \leq ||T-I||_{L^\infty} \leq \max_{x\in D}|y-x| \leq \frac{1}{\lambda}\int_x^y \rho(s)ds,$$

where T is the transportation map between $\widetilde{\nu}$ and ν and $y=Tx$. In high dimensions, it is not clear to us how to bound $W_\infty(\nu,\widetilde{\nu})$ in terms of certain integral of the density. This is to be investigated in our future work.

Acknowledgements. The research was partially supported by KI-Net NSF RNMS11-07444 and NSF DMS-1812573. The authors would like to thank Johannes Wiesel for suggesting the use of the Dvoretzky-Kiefer-Wolfowitz inequality to prove Theorem 1.1, which simplifies an earlier version of the proof of the theorem.

REFERENCES

Ajtai1984 [1] M. Ajtai, J. Komlós, and G. Tusnády, *On optimal matchings*, Combinatorica **4** (1984), no. 4, 259–264.

rthe2013combinatorial [2] Franck Barthe and Charles Bordenave, *Combinatorial optimization over two random point sets*, Séminaire de Probabilités XLV, Springer, 2013, pp. 483–535.

Bernstein1927 [3] S.N. Bernstein, *The theory of probabilities*, Gastehizdat Publishing House,Moscow, 1946.

boissard2011simple [4] Emmanuel Boissard et al., *Simple bounds for the convergence of empirical and occupation measures in 1-wasserstein distance*, Electronic Journal of Probability **16** (2011), 2296–2333.

boissard2014mean [5] Emmanuel Boissard, Thibaut Le Gouic, et al., *On the mean speed of convergence of empirical and occupation measures in wasserstein distance*, Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, vol. 50, Institut Henri Poincaré, 2014, pp. 539–563.

olley2007quantitative [6] François Bolley, Arnaud Guillin, and Cédric Villani, *Quantitative concentration inequalities for empirical measures on non-compact spaces*, Probability Theory and Related Fields **137** (2007), no. 3-4, 541–593.

ampion2008wasserstein [7] Thierry Champion, Luigi De Pascale, and Petri Juutinen, *The ∞-wasserstein distance: Local solutions and existence of optimal transport maps*, SIAM Journal on Mathematical Analysis **40** (2008), no. 1, 1–20.

Chernoff_1952 [8] Herman Chernoff, *A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations*, The Annals of Mathematical Statistics **23** (1952), no. 4, 493–507.

d2006using [9] Michele d'Amico, Patrizio Frosini, and Claudia Landi, *Using matching distance in size theory: A survey*, International Journal of Imaging Systems and Technology **16** (2006), no. 5, 154–161.

davis2018 [10] Erik Davis and Sunder Sethuraman, *Consistency of modularity clustering on random geometric graphs*, Ann. Appl. Probab. **28** (2018), no. 4, 2003–2062.

dudley1969speed [11] RM Dudley, *The speed of mean glivenko-cantelli convergence*, The Annals of Mathematical Statistics **40** (1969), no. 1, 40–50.

voretzky1956asymptotic [12] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz, *Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator*, The Annals of Mathematical Statistics (1956), 642–669.

fournier2015rate [13] Nicolas Fournier and Arnaud Guillin, *On the rate of convergence in wasserstein distance of the empirical measure*, Probability Theory and Related Fields **162** (2015), no. 3-4, 707–738.

gibbs2002choosing [14] Alison L Gibbs and Francis Edward Su, *On choosing and bounding probability metrics*, International statistical review **70** (2002), no. 3, 419–435.

Leighton1989 [15] T. Leighton and P. Shor, *Tight bounds for minimax grid matching with applications to the average case analysis of algorithms*, Combinatorica **9** (1989), no. 2, 161–187.

shor1991 [16] P. W. Shor and J. E. Yukich, *Minimax grid matching and empirical measures*, Ann. Probab. **19** (1991), no. 3, 1338–1348.

chebichef1867 [17] P. Tchebichef, *Des valeurs moyennes*, Journal de mathmatiques pures et appliques **12** (1867), no. 2, 177–184.

rillos2016variational [18] Nicolás García Trillos and Dejan Slepčev, *A variational approach to the consistency of spectral clustering*, Applied and Computational Harmonic Analysis (2016).

trillos2014rate [19] Nicolás Garcia Trillos and Dejan Slepčev, *On the rate of convergence of empirical measures in ∞-transportation distance*, Canadian Journal of Mathematics **67** (2014), 1358.

van1996weak [20] Aad W Van Der Vaart and Jon A Wellner, *Weak convergence*, Weak convergence and empirical processes, Springer, 1996, pp. 16–28.

villani2003topics [21] Cédric Villani, *Topics in optimal transportation*, no. 58, American Mathematical Soc., 2003.

von2007tutorial [22] Ulrike Von Luxburg, *A tutorial on spectral clustering*, Statistics and computing **17** (2007), no. 4, 395–416.