



HAL
open science

Strong equivalence between metrics of Wasserstein type

Erhan Bayraktar, Gaoyue Guo

► **To cite this version:**

Erhan Bayraktar, Gaoyue Guo. Strong equivalence between metrics of Wasserstein type. Electronic Communications in Probability, 2021, 26, 10.1214/21-ecp383 . hal-04584954

HAL Id: hal-04584954

<https://hal.science/hal-04584954>

Submitted on 23 May 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Strong equivalence between metrics of Wasserstein type^{*†}

Erhan Bayraktar[‡] Gaoyue Guo[§]

Abstract

The sliced Wasserstein metric \mathcal{W}_p and more recently max-sliced Wasserstein metric $\overline{\mathcal{W}}_p$ have attracted abundant attention in data sciences and machine learning due to their advantages to tackle the curse of dimensionality, see e.g. [15], [6]. A question of particular importance is the strong equivalence between these projected Wasserstein metrics and the (classical) Wasserstein metric \mathcal{W}_p . Recently, Paty and Cuturi have proved in [14] the strong equivalence of $\overline{\mathcal{W}}_2$ and \mathcal{W}_2 . We show that the strong equivalence also holds for $p = 1$, while the sliced Wasserstein metric does not share this nice property.

Keywords: optimal transport; Wasserstein metric; max-sliced Wasserstein metric; duality.

MSC2020 subject classifications: 90C25.

Submitted to ECP on December 17, 2019, final version accepted on February 19, 2021.

1 Introduction

The Wasserstein metric arising in the optimal transport theory forms a distance function between probability measures. In mathematical language, the Wasserstein distance of order $p \geq 1$ between probability measures μ and ν on \mathbb{R}^d is defined as

$$\mathcal{W}_p(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu, \nu)} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^p \gamma(dx, dy) \right)^{1/p},$$

where $\Gamma(\mu, \nu)$ is the set of probability measures γ on $\mathbb{R}^d \times \mathbb{R}^d$ having the marginal distributions μ and ν . Theoretical advances in the last fifty years characterize existence, uniqueness, representation and smoothness properties of optimizers for $\mathcal{W}_p(\mu, \nu)$ under different settings and compute $\mathcal{W}_p(\mu, \nu)$ by adopting tools and methods in PDE, linear programming and computational geometry, see e.g. [16], [19], and applications are various throughout most of the applied sciences including economics, geography and biomedical sciences, see e.g. [17], [18]. Recently, it has attracted abundant attention in data sciences and machine learning due to its theoretical properties and applications

*E. Bayraktar is supported in part by the Susan M. Smith Professorship and by the National Science Foundation.

†G. Guo is grateful for the support of CentraleSupélec, and in addition, to the University of Michigan and AMS Simons Travel Grant.

‡University of Michigan, United States of America.

E-mail: erhan@umich.edu, Web-page: https://sites.lsa.umich.edu/erhan/

§Université Paris-Saclay, France.

E-mail: gaoyue.guo@centralesupelec.fr

on many domains, see e.g. Wasserstein GAN (Generative Adversarial Network) in [2]. While the Wasserstein metric brings new perspectives and principled ways to formalize problems, the related methods usually suffer from high computational complexity as evaluating Wasserstein distance for $d \gg 1$ is numerically intractable in general. This important computational burden is a major limiting factor in the application of Wasserstein metric to large-scale data analysis. An appealing path to overcome the curse of dimensionality is the recently introduced sliced Wasserstein metric \mathcal{W}_p , which is based on the average Wasserstein distance of the projections of two distributions, see e.g. [15], [3], [18]. Very recently, in order to reduce the projection complexity of the sliced Wasserstein, [6] introduced the so-called max-Wasserstein metrics, which we will denote by $\overline{\mathcal{W}}_p$, as a fix. The same paper also points out that both of these projected versions of the Wasserstein distance enjoy the so-called generalizability over the Wasserstein metric. For further recent results please refer to [13], [14].

Paty and Cuturi showed in [14] that the max-sliced distance $\overline{\mathcal{W}}_2$ is strongly equivalent to \mathcal{W}_2 . This paper aims to prove this result for $p = 1$. The proof of our result is based on the dual formulation of \mathcal{W}_1 , constructing a tailor-made topology τ on the space of Lipschitz functions on \mathbb{R}^d , and some functional analytic arguments. This is reminiscent of the universal approximation result in e.g. [5], i.e. any arbitrary Lipschitz function can be recovered from functional evaluation of projections. Although in the same spirit, our result here is different. In Theorem 3.5, we prove there exists $C_d > 0$ such that the collection of functions as below

$$\sum_{1 \leq k \leq n} a_k f_k(v_k \cdot x) \tag{1.1}$$

is dense, endowed with τ , in the absorbing and convex set of 1-Lipschitz functions $\text{Lip}_1(\mathbb{R}^d)$, where $n \in \mathbb{N}$, $a_k \geq 0$, $v_k \in \mathbb{S}^{d-1}$, $f_k \in \text{Lip}_1(\mathbb{R})$ for $1 \leq k \leq n$ and $\sum_{1 \leq k \leq n} a_k \leq C_d$. Roughly speaking, any 1-Lipschitz function on \mathbb{R}^d can be approximated by a sequence of C_d -Lipschitz functions of form (1.1).

We show further that the strong equivalence is not shared by the sliced Wasserstein metric using the recent results of [1], hence promoting the max-sliced metric over the sliced one.

The structure of the rest of the paper is simple. In the next section, after introducing some preliminaries we will give our main results in Theorem 2.3. Section 3, on the other hand, is devoted to the proof of these results. A technical lemma is presented in the Appendix.

2 Main results

2.1 Preliminaries on the Wasserstein metric

We start by reviewing the preliminary concepts and formulations needed to introduce the main results. For $p \geq 1$, let $\mathcal{P}_p(\mathbb{R}^d)$ be the set of probability measures on \mathbb{R}^d admitting finite p^{th} moment, i.e. $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ if and only if

$$M_p(\mu) := \left(\int_{\mathbb{R}^d} |x|^p \mu(dx) \right)^{1/p} < \infty. \tag{2.1}$$

For $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, denote by $\Gamma(\mu, \nu) \subset \mathcal{P}_p(\mathbb{R}^d \times \mathbb{R}^d)$ the collection of probability measures γ on $\mathbb{R}^d \times \mathbb{R}^d$, also known as couplings or transport plans, such that

$$\gamma[E \times \mathbb{R}^d] = \mu[E] \text{ and } \gamma[\mathbb{R}^d \times E] = \nu[E], \text{ for all measurable sets } E \subset \mathbb{R}^d.$$

The Wasserstein metric of order p is a distance function $\mathcal{W}_p : \mathcal{P}_p(\mathbb{R}^d) \times \mathcal{P}_p(\mathbb{R}^d) \rightarrow \mathbb{R}_+$ defined by

$$\mathcal{W}_p(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu, \nu)} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^p \gamma(dx, dy) \right)^{1/p}, \quad \text{for all } \mu, \nu \in \mathcal{P}_p(\mathbb{R}^d). \quad (2.2)$$

It is known that $\mathcal{P}_p(\mathbb{R}^d)$ endowed with \mathcal{W}_p is a Polish space, i.e. separable completely metrizable space, see e.g. Theorem 6.18 of [19]. In particular, an explicit expression of $\mathcal{W}_p(\mu, \nu)$ is given for $d = 1$:

$$\mathcal{W}_p(\mu, \nu) = \left(\int_0^1 |F_\mu^{-1}(t) - F_\nu^{-1}(t)|^p dt \right)^{1/p}, \quad (2.3)$$

where $F_\mu^{-1}(t) := \inf\{x \in \mathbb{R} : \mu[(-\infty, x]] > t\}$ and $F_\nu^{-1}(t) := \inf\{x \in \mathbb{R} : \nu[(-\infty, x]] > t\}$, see e.g. Chapter 2 of [18]. We note also that \mathcal{W}_p is depending on d , i.e. $\mathcal{W}_p \equiv \mathcal{W}_{p,d}$. Nevertheless, we will not emphasize this dependency in the rest and write simply \mathcal{W}_p without any danger of confusion.

2.2 Projected Wasserstein metrics

While approaches based on the Wasserstein metric have been successful in several complex tasks, estimating the Wasserstein distance often suffers from the curse of dimensionality from the complexity/algorithmic perspective. To tackle the issue of complexity, a sliced version of the Wasserstein distance was studied and employed, which only requires estimating distances of projected uni-dimensional distributions and is, therefore, more efficient, see e.g. [3], [6], [13].

Let $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ be the unit sphere, i.e. $\mathbb{S}^{d-1} := \{v \in \mathbb{R}^d : |v| = 1\}$. For $\mu \in \mathcal{P}_p(\mathbb{R}^d)$ and $v \in \mathbb{S}^{d-1}$, set $\mu_v := \mu \circ v_*^{-1}$ to be the image measure of μ by v_* , where $v_* : \mathbb{R}^d \rightarrow \mathbb{R}$ is the map defined by $v_*(x) := v \cdot x$. Then $\mu_v \in \mathcal{P}_p(\mathbb{R})$ since

$$\int_{\mathbb{R}} |x|^p \mu_v(dx) = \int_{\mathbb{R}^d} |v \cdot x|^p \mu(dx) \leq \int_{\mathbb{R}^d} |x|^p \mu(dx).$$

Hence, we may define the sliced Wasserstein metric \mathcal{W}_p and max-sliced Wasserstein metric $\overline{\mathcal{W}}_p$ as follows.

Definition 2.1. For $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, set

$$\mathcal{W}_p^s(\mu, \nu) := \left(\oint_{\mathbb{S}^{d-1}} \mathcal{W}_p(\mu_v, \nu_v)^p dv \right)^{1/p} \quad \text{and} \quad \overline{\mathcal{W}}_p(\mu, \nu) := \sup_{v \in \mathbb{S}^{d-1}} \mathcal{W}_p(\mu_v, \nu_v),$$

where \oint denotes the surface integral over \mathbb{S}^{d-1} .

For $d = 1$, one has by definition $\mathcal{W}_p(\mu, \nu) = 2^{-1/p} \mathcal{W}_p^s(\mu, \nu) = \overline{\mathcal{W}}_p(\mu, \nu)$. Further, Proposition 2.2 ensures that $\mathcal{W}_p^s, \overline{\mathcal{W}}_p$ are well defined metrics on $\mathcal{P}_p(\mathbb{R}^d)$.

Proposition 2.2. (i) Fix $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$. The maps $\mathbb{S}^{d-1} \ni v \mapsto \mu_v \in \mathcal{P}_p(\mathbb{R})$ and $\mathbb{S}^{d-1} \ni v \mapsto \mathcal{W}_p(\mu_v, \nu_v) \in \mathbb{R}_+$ are both Lipschitz, with respectively Lipschitz constants $M_p(\mu)$ and $M_p(\mu) + M_p(\nu)$. In particular, the supremum over \mathbb{S}^{d-1} can be attained and $\overline{\mathcal{W}}_p(\mu, \nu) = \max_{v \in \mathbb{S}^{d-1}} \mathcal{W}_p(\mu_v, \nu_v)$.

(ii) \mathcal{W}_p^s and $\overline{\mathcal{W}}_p$ form two distance functions on $\mathcal{P}_p(\mathbb{R}^d)$.

Proof. (i) For all $u, v \in \mathbb{S}^{d-1}$, let $u_* \otimes v_* : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^2$ be the map defined by $u_* \otimes v_*(x, y) := (u \cdot x, v \cdot y)$. Taking $\gamma(dx, dy) := \mu(dx) \otimes \delta_x(dy) \in \Gamma(\mu, \mu)$, one has $\gamma_{u,v} :=$

$\gamma \circ (\mathbf{u}_* \otimes \mathbf{v}_*)^{-1} \in \Gamma(\mu_u, \mu_v)$ and thus

$$\begin{aligned} \mathcal{W}_p(\mu_u, \mu_v) &\leq \left(\int_{\mathbb{R}^2} |x - y|^p \gamma_{u,v}(dx, dy) \right)^{1/p} \\ &= \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} |\mathbf{u} \cdot \mathbf{x} - \mathbf{v} \cdot \mathbf{y}|^p \gamma(dx, dy) \right)^{1/p} \leq |\mathbf{u} - \mathbf{v}| M_p(\mu). \end{aligned}$$

Further, the triangle inequality yields $|\mathcal{W}_p(\mu_u, \nu_u) - \mathcal{W}_p(\mu_v, \nu_v)| \leq |\mathcal{W}_p(\mu_u, \mu_v)| + |\mathcal{W}_p(\nu_u, \nu_v)| \leq |\mathbf{u} - \mathbf{v}|(M_p(\mu) + M_p(\nu))$, which yields the Lipschitz continuity and further $\overline{\mathcal{W}}_p(\mu, \nu) = \max_{\mathbf{v} \in \mathbb{S}^{d-1}} \mathcal{W}_p(\mu_v, \nu_v)$.

(ii) The symmetry and sub-additivity are trivial by definition. Denote by A_d the surface area of \mathbb{S}^{d-1} , i.e.

$$A_d := \oint_{\mathbb{S}^{d-1}} d\mathbf{v} = \frac{2\pi^{d/2}}{\Gamma(d/2)}, \tag{2.4}$$

where $\Gamma : \mathbb{R} \rightarrow \mathbb{R}$ is the Gamma function given as

$$\Gamma(x) := \int_0^\infty t^{x-1} e^{-t} dt.$$

Then one has $\mathcal{W}_p(\mu, \nu) \leq A_d^{1/p} \overline{\mathcal{W}}_p(\mu, \nu)$ for all $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ and it suffices to show the identity of indiscernibles for \mathcal{W}_p . Let $\mathcal{W}_p(\mu, \nu) = 0$, which implies by (i) that $\mu_v = \nu_v$ for all $\mathbf{v} \in \mathbb{S}^{d-1}$. Consider the characteristic functions of μ, ν defined by

$$\tilde{\mu}(\mathbf{z}) := \int_{\mathbb{R}^d} e^{i\mathbf{z} \cdot \mathbf{x}} \mu(d\mathbf{x}) \quad \text{and} \quad \tilde{\nu}(\mathbf{z}) := \int_{\mathbb{R}^d} e^{i\mathbf{z} \cdot \mathbf{x}} \nu(d\mathbf{x}).$$

With $r := |\mathbf{z}|$ and $\mathbf{v} := \mathbf{z}/r$, it follows that

$$\int_{\mathbb{R}^d} e^{i\mathbf{z} \cdot \mathbf{x}} \mu(d\mathbf{x}) = \int_{\mathbb{R}} e^{irx} \mu_v(dx) = \int_{\mathbb{R}} e^{irx} \nu_v(dx) = \int_{\mathbb{R}^d} e^{i\mathbf{z} \cdot \mathbf{x}} \nu(d\mathbf{x}),$$

which implies $\tilde{\mu}(\mathbf{z}) = \tilde{\nu}(\mathbf{z})$ for all $\mathbf{z} \in \mathbb{R}^d$ and finally $\mu = \nu$. □

2.3 Main results

Given the active theoretical interest of Wasserstein metric, as well as its importance for applications in practice, the investigation of \mathcal{W}_p and $\overline{\mathcal{W}}_p$ is gaining popularity in machine learning, with several applications to data sciences. A question of particular importance is the equivalence between $\mathcal{W}_p, \overline{\mathcal{W}}_p$ and \mathcal{W}_p . Recently, Paty and Cuturi proved in [14] the strong equivalence of $\overline{\mathcal{W}}_2$ and \mathcal{W}_2 . Namely,

$$\overline{\mathcal{W}}_2(\mu, \nu) \leq \mathcal{W}_2(\mu, \nu) \leq \sqrt{d} \overline{\mathcal{W}}_2(\mu, \nu), \quad \text{for all } \mu, \nu \in \mathcal{P}_2(\mathbb{R}^d).$$

In this paper, we show the (topological) equivalence between $\mathcal{W}_p, \overline{\mathcal{W}}_p$ and \mathcal{W}_p as well as the strong equivalence between $\overline{\mathcal{W}}_1$ and \mathcal{W}_1 , which are summarized in Theorem 2.3 below.

Theorem 2.3. (i) $\mathcal{W}_p, \overline{\mathcal{W}}_p$ and \mathcal{W}_p are equivalent for all $p \geq 1$, i.e.

$$\lim_{n \rightarrow \infty} \mathcal{W}_p(\mu^n, \mu) = 0 \iff \lim_{n \rightarrow \infty} \overline{\mathcal{W}}_p(\mu^n, \mu) = 0 \iff \lim_{n \rightarrow \infty} \mathcal{W}_p(\mu^n, \mu) = 0$$

for any sequence $(\mu^n)_{n \geq 1} \subset \mathcal{P}_p(\mathbb{R}^d)$ and $\mu \in \mathcal{P}_p(\mathbb{R}^d)$.

(ii) $\overline{\mathcal{W}}_1$ and \mathcal{W}_1 are strongly equivalent for all $d \geq 1$, i.e. there exists $C_d \geq 1$ such that

$$\overline{\mathcal{W}}_1(\mu, \nu) \leq \mathcal{W}_1(\mu, \nu) \leq C_d \overline{\mathcal{W}}_1(\mu, \nu), \quad \text{for all } \mu, \nu \in \mathcal{P}_1(\mathbb{R}^d). \tag{2.5}$$

(iii) \mathcal{W}_1 and \mathcal{W}_1 are not strongly equivalent for all $d \geq 2$.

Remark 2.4.¹ For every $R > 0$, denote by $\mathcal{P}(B_R) \subset \mathcal{P}_1(\mathbb{R}^d)$ the subset of probability measures supported in $B_R := \{x \in \mathbb{R}^d : |x| \leq R\}$. Then one must have $0 < C_d^* \leq C_d$ such that

$$\mathcal{W}_1(\mu, \nu) \leq C_d^* \overline{\mathcal{W}}_1(\mu, \nu), \quad \text{for all } \mu, \nu \in \mathcal{P}(B_1).$$

Let $\mu \in \mathcal{P}(B_R)$ for some $R > 0$ and $X \sim \mu$. Denote by μ^R the distribution of X/R , then clearly $\mu^R \in \mathcal{P}(B_1)$. Further, for $\mu, \nu \in \mathcal{P}(B_R)$, one has $\mathcal{W}_1(\mu, \nu) = R\mathcal{W}_1(\mu^R, \nu^R)$ and $\overline{\mathcal{W}}_1(\mu, \nu) = R\overline{\mathcal{W}}_1(\mu^R, \nu^R)$, which implies

$$\mathcal{W}_1(\mu, \nu) \leq C_d^* \overline{\mathcal{W}}_1(\mu, \nu), \quad \text{for all } \mu, \nu \in \bigcup_{R>0} \mathcal{P}(B_R).$$

and further $\mathcal{W}_1 \leq C_d^* \overline{\mathcal{W}}_1$ on $\mathcal{P}_1(\mathbb{R}^d)$ as $\bigcup_{R>0} \mathcal{P}(B_R)$ is dense in $\mathcal{P}_1(\mathbb{R}^d)$ under \mathcal{W}_1 and $\overline{\mathcal{W}}_1$.

3 Proof of the main results

3.1 Proof of Theorem 2.3 (i)

Given $u, v \in \mathbb{S}^{d-1}$, let $u_* \otimes v_* : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^2$ be the map defined by

$$u_* \otimes v_*(x, y) := (u \cdot x, v \cdot y). \tag{3.1}$$

For $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, let $\gamma \in \Gamma(\mu, \nu)$ be an optimizer for $\mathcal{W}_p(\mu, \nu)$. Then, by definition, $\gamma_{u,v} := \gamma \circ (u_* \otimes v_*)^{-1} \in \Gamma(\mu_u, \nu_v)$. Taking in particular $u = v$, one has

$$\mathcal{W}_p(\mu_v, \nu_v)^p \leq \int_{\mathbb{R}^2} |x - y|^p \gamma_{v,v}(dx, dy) = \int_{\mathbb{R}^d \times \mathbb{R}^d} |v \cdot x - v \cdot y|^p \gamma(dx, dy) \leq \mathcal{W}_p(\mu, \nu)^p,$$

which yields the trivial inequality as follows:

$$\frac{1}{A_d^{1/p}} \mathcal{W}_p(\mu, \nu) \leq \overline{\mathcal{W}}_p(\mu, \nu) \leq \mathcal{W}_p(\mu, \nu), \tag{3.2}$$

where A_d is defined in (2.4). Therefore, it remains to show the equivalence between \mathcal{W}_p and $\overline{\mathcal{W}}_p$.

For the sake of presentation, we use the following notation for the fact that random variable X is distributed according to probability measure μ :

$$X \sim \mu \quad \text{or} \quad \mathcal{L}(X) = \mu,$$

where $\mathcal{L}(X)$ denotes the law of X .

of Theorem 2.3 (i). It suffices to prove that $\lim_{n \rightarrow \infty} \mathcal{W}_p(\mu^n, \mu) = 0$ implies $\lim_{n \rightarrow \infty} \overline{\mathcal{W}}_p(\mu^n, \mu) = 0$. We proceed as follows.

Step 1. For each $n \geq 1$, let X^n be a random variable with $X^n \sim \mu^n$, and by definition, $v \cdot X^n \sim \mu_v^n$ holds for all $v \in \mathbb{S}^{d-1}$. As

$$\lim_{n \rightarrow \infty} \oint_{\mathbb{S}^{d-1}} \mathcal{W}_p(\mu_v^n, \mu_v)^p d\nu = \lim_{n \rightarrow \infty} \mathcal{W}_p(\mu_v^n, \mu_v)^p = 0,$$

the functions $\mathbb{S}^{d-1} \ni v \mapsto \mathcal{W}_p(\mu_v^n, \mu_v)^p \in \mathbb{R}_+$ converge in measure to zero. Namely, for almost every $v \in \mathbb{S}^{d-1}$, $\lim_{n \rightarrow \infty} \mathcal{W}_p(\mu_v^n, \mu_v) = 0$. We can further conclude that the

¹This observation is suggested the anonymous referee and provides a tractable schema to approximate C_d^* by solving an optimization problem over the compact subset $\mathcal{P}_1(B_1)$.

sequence $(|v \cdot X^n|^p)_{n \geq 1}$ is uniformly integrable. Pick a finite set $\{v_1, \dots, v_I\} \subset \mathbb{S}^{d-1}$ such that

$$\lim_{n \rightarrow \infty} \mathcal{W}_p(\mu_{v_i}^n, \mu_{v_i}) = 0 \text{ for } 1 \leq i \leq I \quad \text{and} \quad |x|/2 \leq \max_{1 \leq i \leq I} |v_i \cdot x| \text{ for all } x \in \mathbb{R}^d.$$

Hence, $|X^n|^p \leq 2^p \sum_{i=1}^I |v_i \cdot X^n|^p$ yields the uniform integrability of $(|X^n|^p)_{n \geq 1}$ and in particular

$$\sup_{n \geq 1} M_p(\mu^n) = \sup_{n \geq 1} \left(\int_{\mathbb{R}^d} |x|^p \mu^n(dx) \right)^{1/p} = \sup_{n \geq 1} (\mathbb{E}[|X^n|^p])^{1/p} =: C < \infty.$$

In view of the proof of Proposition 2.2, the maps $\mathbb{S}^{d-1} \ni v \mapsto \mathcal{W}_p(\mu_v^n, \mu_v)^p \in \mathbb{R}_+$ are equi-Lipschitz with a uniform Lipschitz constant $C + M_p(\mu)$, and thus

$$\lim_{n \rightarrow \infty} \mathcal{W}_p(\mu_v^n, \mu_v) = 0, \quad \text{for all } v \in \mathbb{S}^{d-1}. \tag{3.3}$$

Step 2. Consider the characteristic function of μ given by

$$\tilde{\mu}(z) := \int_{\mathbb{R}^d} e^{iz \cdot x} \mu(dx), \quad \text{for all } z \in \mathbb{R}^d.$$

Define similarly $\tilde{\mu}^n$ for all $n \geq 1$. For every $z \in \mathbb{R}^d$, with $r := |z|$ and $v := z/r$, it holds

$$\lim_{n \rightarrow \infty} \int_{\mathbb{R}^d} e^{iz \cdot x} \mu^n(dx) = \lim_{n \rightarrow \infty} \int_{\mathbb{R}^d} e^{irx} \mu_v^n(dx) = \int_{\mathbb{R}^d} e^{irx} \mu_v(dx) = \int_{\mathbb{R}^d} e^{iz \cdot x} \mu(dx),$$

where the second equality follows from (3.3). We conclude thus $(\mu^n)_{n \geq 1}$ converges weakly to μ .

Step 3. Using the Skorokhod representation theorem, we may assume without loss of generality that the sequence $(X^n)_{n \geq 1}$ converges almost surely. Denote by $X \sim \mu$ its limit. Combining with the uniform integrability of $(|X^n|^p)_{n \geq 1}$, one has

$$\lim_{n \rightarrow \infty} \mathcal{W}_p(\mu^n, \mu)^p \leq \lim_{n \rightarrow \infty} \mathbb{E}[|X^n - X|^p] = 0,$$

which fulfils the proof. □

3.2 Proof of Theorem 2.3 (ii)

Our proof is based on the dual formulation of \mathcal{W}_1 and inspired by the proof of the universal approximation theorem. Let $L^0(\mathbb{R}^d)$ be space of Lebesgue measurable functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$ and $\text{Lip}(\mathbb{R}^d) \subset L^0(\mathbb{R}^d)$ be the subspace consisting of Lipschitz functions vanishing at the origin, i.e. $f \in \text{Lip}(\mathbb{R}^d)$ if and only if $f(0) = 0$ and

$$\|f\|_{\text{lip}} := \sup_{x \neq y \in \mathbb{R}^d} \frac{|f(x) - f(y)|}{|x - y|} < \infty.$$

For each $L > 0$, denote by $\text{Lip}_L(\mathbb{R}^d) \subset \text{Lip}(\mathbb{R}^d)$ the subset of functions f with $\|f\|_{\text{lip}} \leq L$. Then it follows by Kantorovich's duality that, see e.g. Remark 6.5 in [19],

$$\mathcal{W}_1(\mu, \nu) = \sup_{f \in \text{Lip}_1(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} f(x) \mu(dx) - \int_{\mathbb{R}^d} f(x) \nu(dx) \right\}, \quad \text{for all } \mu, \nu \in \mathcal{P}_1(\mathbb{R}^d) \tag{3.4}$$

In what follows, (3.4) will be used in the proof of Theorem 2.3. It is known from [12], $\|\cdot\|_{\text{lip}}$ defines a norm on $\text{Lip}(\mathbb{R}^d)$ and $(\text{Lip}(\mathbb{R}^d), \|\cdot\|_{\text{lip}})$ is a Banach space. Next we endow

$\text{Lip}(\mathbb{R}^d)$ with an alternative topology. Set

$$\begin{aligned} L^1(\mathbb{R}^d) &:= \left\{ f \in L^0(\mathbb{R}^d) : \|f\|_1 := \int_{\mathbb{R}^d} |f(x)| dx < \infty \right\} \\ L^\infty(\mathbb{R}^d) &:= \left\{ f \in L^0(\mathbb{R}^d) : \|f\|_\infty := \text{ess sup}_{x \in \mathbb{R}^d} |f(x)| < \infty \right\}, \end{aligned}$$

then $(L^1(\mathbb{R}^d), \|\cdot\|_1)$ and $(L^\infty(\mathbb{R}^d), \|\cdot\|_\infty)$ are both Banach spaces. Denote further by $L^1(\mathbb{R}^d)^d$ (resp. $L^\infty(\mathbb{R}^d)^d$) the d -product of $L^1(\mathbb{R}^d)$ (resp. $L^\infty(\mathbb{R}^d)^d$). Note in particular that every $f \in \text{Lip}(\mathbb{R}^d)$ is a.e. differentiable and $\nabla f \in L^\infty(\mathbb{R}^d)^d$ with $\|\nabla f\|_\infty = \|f\|_{\text{lip}}$, see e.g. Exercise 6.14 of [11]. Finally we define $L^1(\mathbb{R}^d, (1 + |x|)dx) \subset L^1(\mathbb{R}^d)$ by

$$L^1(\mathbb{R}^d, (1 + |x|)dx) := \left\{ f \in L^1(\mathbb{R}^d) : \int_{\mathbb{R}^d} (1 + |x|)|f(x)| dx < \infty \right\}.$$

Now we are ready to introduce the topology, denoted by τ , on $\text{Lip}(\mathbb{R}^d)$ for which $(\text{Lip}(\mathbb{R}^d), \tau)$ is a locally convex space. Let τ^o be a collection of subsets $\mathcal{O}_{(u,w)}(f; \varepsilon) \subset \text{Lip}(\mathbb{R}^d)$ defined as follows:

$$\mathcal{O}_{(u,w)}(f; \varepsilon) := \left\{ g \in \text{Lip}(\mathbb{R}^d) : \left| \int_{\mathbb{R}^d} [(g - f)(x)u(x) + \nabla(g - f)(x) \cdot w(x)] dx \right| < \varepsilon \right\},$$

where $f \in \text{Lip}(\mathbb{R}^d)$, $\varepsilon > 0$ and $(u, w) \in L^1(\mathbb{R}^d, (1 + |x|)dx) \times L^1(\mathbb{R}^d)^d$. Then we define τ to be the topology generated by τ^o . $(\text{Lip}(\mathbb{R}^d), \tau)$ is a locally convex space, as the origin has a local base of absolutely convex absorbent sets, see e.g. Proposition 1.15 in Chapter 4 of [4]. Under the topology τ , $(f_n)_{n \geq 1} \subset \text{Lip}(\mathbb{R}^d)$ converges to $f \in \text{Lip}(\mathbb{R}^d)$ if and only if

$$\lim_{n \rightarrow \infty} \int_{\mathbb{R}^d} [(f_n - f)(x)u(x) + \nabla(f_n - f)(x) \cdot w(x)] dx = 0$$

holds for all $(u, w) \in L^1(\mathbb{R}^d, (1 + |x|)dx) \times L^1(\mathbb{R}^d)^d$.

Remark 3.1. (i) In view of [9], $\text{Lip}(\mathbb{R}^d)$ is the dual space of some closed quotient space of $L^1(\mathbb{R}^d)^d$ and τ turns to be the weak* topology. Hence, $(\text{Lip}(\mathbb{R}^d), \tau)$ is not metrizable and thus not a Fréchet space as this quotient space is infinite dimensional, see e.g. Chapter 5 of [4]. However, Proposition 3.4 shows that τ , restricted to $\text{Lip}_L(\mathbb{R}^d) \subset \text{Lip}(\mathbb{R}^d)$ for any $L > 0$, is metrizable.

(ii) For any sequence $(g_n)_{n \geq 1} \subset \text{Lip}(\mathbb{R}^d)$ with $\lim_{n \rightarrow \infty} \|g_n\|_{\text{lip}} = 0$, one has for any $(u, w) \in L^1(\mathbb{R}^d, (1 + |x|)dx) \times L^1(\mathbb{R}^d)^d$

$$\begin{aligned} \lim_{n \rightarrow \infty} \left| \int_{\mathbb{R}^d} g_n(x)u(x) dx \right| &= \lim_{n \rightarrow \infty} \left| \int_{\mathbb{R}^d} \left(\int_0^1 x \cdot \nabla g_n(tx) dt \right) u(x) dx \right| \\ &\leq \lim_{n \rightarrow \infty} \|g_n\|_{\text{lip}} \int_{\mathbb{R}^d} |x| |u(x)| dx = 0 \end{aligned}$$

and

$$\lim_{n \rightarrow \infty} \left| \int_{\mathbb{R}^d} \nabla g_n(x) \cdot w(x) dx \right| \leq \lim_{n \rightarrow \infty} \|g_n\|_{\text{lip}} \int_{\mathbb{R}^d} |w(x)| dx = 0.$$

Therefore, the topology under $\|\cdot\|_{\text{lip}}$ is strictly stronger than τ .

The lemma below characterizes the space of τ -continuous linear functions on $\text{Lip}(\mathbb{R}^d)$.

Lemma 3.2. Any τ -continuous linear functions $T : \text{Lip}(\mathbb{R}^d) \rightarrow \mathbb{R}$ must be of the form

$$T(f) = \int_{\mathbb{R}^d} [f(x)u(x) + \nabla f(x) \cdot w(x)] dx,$$

for some $(u, w) \in L^1(\mathbb{R}^d, (1 + |x|)dx) \times L^1(\mathbb{R}^d)^d$.

Let $\mathcal{C} \subset \text{Lip}_1(\mathbb{R}^d)$ be the subset of functions f of the form:

$$f(x) = \sum_{1 \leq k \leq n} a_k f_k(\mathbf{v}_k \cdot x),$$

where $n \in \mathbb{N}$, $a_k \in \mathbb{R}_+$, $\mathbf{v}_k \in \mathbb{S}^{d-1}$, $f_k \in \text{Lip}_1(\mathbb{R})$ for $1 \leq k \leq n$ and $\sum_{1 \leq k \leq n} a_k = 1$. Define further $m\mathcal{C} := \{af : |a| \leq m \text{ and } f \in \mathcal{C}\}$ for each $m \geq 1$ and $\mathcal{D} := \bigcup_{m \geq 1} m\mathcal{C}$. We denote by $\overline{\mathcal{D}}$ the τ -closure of \mathcal{D} in $\text{Lip}(\mathbb{R}^d)$. Similarly, we define by $m\overline{\mathcal{C}}$ the τ -closure of $m\mathcal{C}$. Then one has the following proposition.

Proposition 3.3. (i) \mathcal{D} is dense in $\text{Lip}(\mathbb{R}^d)$ with respect to τ , i.e. $\overline{\mathcal{D}} = \text{Lip}(\mathbb{R}^d)$.

(ii) Further, $\text{Lip}(\mathbb{R}^d) = \bigcup_{m \geq 1} m\overline{\mathcal{C}}$ holds.

Proof. (i) If $\overline{\mathcal{D}} \neq \text{Lip}(\mathbb{R}^d)$, then by the Hahn-Banach theorem, see e.g. Corollary 3.15 in Chapter 4 of [4], there exists a non-zero τ -continuous linear function $T : \text{Lip}(\mathbb{R}^d) \rightarrow \mathbb{R}$ such that $T(f) = 0$ for all $f \in \mathcal{D}$, where

$$T(f) = \int_{\mathbb{R}^d} [f(x)u(x) + \nabla f(x) \cdot w(x)] dx, \quad \text{for some } (u, w) \in L^1(\mathbb{R}^d, (1 + |x|)dx) \times L^1(\mathbb{R}^d)^d.$$

Take the bump function $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}_+$ defined by

$$\varphi(x) := \begin{cases} c \exp(1/(|x|^2 - 1)) & \text{if } |x| \leq 1 \\ 0 & \text{otherwise,} \end{cases}$$

where $c > 0$ is chosen such that $\int_{\mathbb{R}^d} \varphi(x) dx = 1$. Define further the sequence $(\varphi_t)_{t>0}$ with $\varphi_t(x) := \varphi(x/t)/t^d$. For every $f \in \mathcal{D}$, one has the convolution $\varphi_t * f \in \mathcal{D}$ and thus

$$\begin{aligned} 0 &= \int_{\mathbb{R}^d} [u(x)(\varphi_t * f)(x) + w(x) \cdot \nabla(\varphi_t * f)(x)] dx \\ &= \int_{\mathbb{R}^d} [u * \varphi_t(x) - \text{div}(w * \varphi_t)(x)] f(x) dx, \end{aligned} \tag{3.5}$$

where the integration by parts can be applied thanks to the convolution. Taking respectively $f(x) = \cos(2\pi z \cdot x)$ and $f(x) = \sin(2\pi z \cdot x)$ for $z \in \mathbb{R}^d$, one deduces that the Fourier transform of $u * \varphi_t - \text{div}(w * \varphi_t)$ is identically equal to zero, i.e.

$$\int_{\mathbb{R}^d} [u * \varphi_t(x) - \text{div}(w * \varphi_t)(x)] e^{-2i\pi z \cdot x} dx = 0, \quad \text{for all } z \in \mathbb{R}^d,$$

and thus $u * \varphi_t - \text{div}(w * \varphi_t) \equiv 0$. Therefore, (3.5) holds for any $f \in \text{Lip}(\mathbb{R}^d)$. Further,

$$\begin{aligned} &\left| \int_{\mathbb{R}^d} [u(x)((\varphi_t * f)(x) - f(x)) + w(x) \cdot (\nabla(\varphi_t * f)(x) - \nabla f(x))] dx \right| \\ &\leq \int_{\mathbb{R}^d} |u(x)| \left(\int_{\mathbb{R}^d} \varphi(y) |f(x - ty) - f(x)| dy \right) dx \\ &\quad + \int_{\mathbb{R}^d} |w(x)| \left| \int_{\mathbb{R}^d} \varphi(y) \nabla f(x - ty) dy - \nabla f(x) \right| dx. \end{aligned}$$

Using the dominated convergence theorem, and the Lebesgue-Besicovitch differentiation theorem, see e.g. page 43 of [7], for the second term, one has $T(f) = 0$ for any $f \in \text{Lip}(\mathbb{R}^d)$, contradicting the fact that T is not null.

(ii) Let $f \in \text{Lip}(\mathbb{R}^d)$. Then there exists a net $(f^\lambda)_\lambda \subset \mathcal{D}$ such that f^λ converges to f under τ . Hence, the continuous linear functions $F_\lambda : L^1(\mathbb{R}^d)^d \rightarrow \mathbb{R}$ defined by

$$F_\lambda(w) := \int_{\mathbb{R}^d} \nabla f^\lambda(x) \cdot w(x) dx$$

are pointwise bounded. By the uniform boundedness principle, it holds $\sup_\lambda \|f^\lambda\|_{\text{lip}} = \sup_\lambda \|\nabla f^\lambda\|_\infty = \sup_\lambda \|F_\lambda\| < \infty$. Thus $f \in m\bar{\mathcal{C}}$ for any $m \geq \sup_\lambda \|f^\lambda\|_{\text{lip}}$. \square

Proposition 3.4. $\bar{\mathcal{C}}$ is closed with respect to the norm $\|\cdot\|_{\text{lip}}$.

Proof. First, we show that the topology τ restricted to $\text{Lip}_1(\mathbb{R}^d) \subset \text{Lip}(\mathbb{R}^d)$ is metrizable. Since $L^1(\mathbb{R}^d, (1 + |\cdot|)dx)$ and $L^1(\mathbb{R}^d)^d$ are separable, we may take two dense subsets $(u_i)_{i \geq 1}$ and $(w_j)_{j \geq 1}$ and define $\rho_{u_i, w_j} : \text{Lip}(\mathbb{R}^d) \times \text{Lip}(\mathbb{R}^d) \rightarrow \mathbb{R}_+$ by

$$\rho_{u_i, w_j}(f, g) := \left| \int_{\mathbb{R}^d} [(f - g)(x)u_i(x) + \nabla(f - g)(x) \cdot w_j(x)] dx \right|.$$

Then by a straightforward verification, the distance $\rho : \text{Lip}(\mathbb{R}^d) \times \text{Lip}(\mathbb{R}^d) \rightarrow \mathbb{R}_+$ given by

$$\rho(f, g) := \sum_{i, j \geq 1} \frac{1}{2^{i+j}} \frac{\rho_{u_i, w_j}(f, g)}{1 + \rho_{u_i, w_j}(f, g)}$$

is consistent with the topology τ restricted on $\text{Lip}_1(\mathbb{R}^d)$. Second, we prove $\bar{\mathcal{C}} \subset \text{Lip}_1(\mathbb{R}^d)$. For any $f \in \bar{\mathcal{C}}$, one has a sequence $(f_n)_{n \geq 1} \subset \mathcal{C}$ converging to f as τ is metrizable on \mathcal{C} . For any $w \in L^1(\mathbb{R}^d)^d$, it follows that

$$\lim_{n \rightarrow \infty} \int_{\mathbb{R}^d} \nabla f_n(x) \cdot w(x) dx = \int_{\mathbb{R}^d} \nabla f(x) \cdot w(x) dx,$$

which means that ∇f_n converges to ∇f under the weak* topology of $L^\infty(\mathbb{R}^d)^d$. Note also, in view of the Banach-Alaoglu theorem, $(\nabla f_n)_{n \geq 1}$ belongs to the unit ball $B_1^\infty \subset L^\infty(\mathbb{R}^d)^d$ which is relatively compact with respect to the weak* topology. Then the uniqueness of the weak* limit yields $\nabla f \in B_1^\infty$, and thus $f \in \text{Lip}_1(\mathbb{R}^d)$. Hence $\bar{\mathcal{C}} \subset \text{Lip}_1(\mathbb{R}^d)$.

Let f be in the closure of $\bar{\mathcal{C}}$ with respect to $\|\cdot\|_{\text{lip}}$. Let $(f^n)_{n \geq 1} \subset \bar{\mathcal{C}}$ satisfying $\lim_{n \rightarrow \infty} \|\nabla(f^n - f)\|_\infty = \lim_{n \rightarrow \infty} \|f^n - f\|_{\text{lip}} = 0$, which implies in particular $\lim_{n \rightarrow \infty} \rho(f^n, f) = 0$ as $|f^n(x) - f(x)| \leq \|f^n - f\|_{\text{lip}}|x|$ for all $x \in \mathbb{R}^d$. For each $n \geq 1$, since $f^n \in \bar{\mathcal{C}}$, there exists $g^n \in \mathcal{C}$ such that $\rho(g^n, f^n) \leq 1/n$. Then $\lim_{n \rightarrow \infty} \rho(g^n, f) \leq \lim_{n \rightarrow \infty} (\rho(g^n, f^n) + \rho(f^n, f)) = 0$, which concludes the proof. \square

Theorem 3.5. There exists a constant $C_d > 0$ such that $\text{Lip}_1(\mathbb{R}^d) \subset C_d \bar{\mathcal{C}}$.

Proof. In view of Propositions 3.3 and 3.4, $\text{Lip}(\mathbb{R}^d) = \bigcup_{m \geq 1} m\bar{\mathcal{C}}$ and $m\bar{\mathcal{C}}$ is closed with respect to $\|\cdot\|_{\text{lip}}$ for each $m \geq 1$. Now it follows from Baire's theorem that there must exist $m^* \geq 1$ such that $m^*\bar{\mathcal{C}}$ has non-empty interior, i.e.

$$\{f \in \text{Lip}(\mathbb{R}^d) : \|f - f^*\|_{\text{lip}} < \varepsilon^*\} \subset m^*\bar{\mathcal{C}}, \text{ for some } f^* \in \text{Lip}(\mathbb{R}^d) \text{ and } \varepsilon^* > 0.$$

By the proof of Proposition 3.3, there exists $m_0 \geq 1$ such that $f^* \in m_0\bar{\mathcal{C}}$. Thus one has

$$\{f \in \text{Lip}(\mathbb{R}^d) : \|f\|_{\text{lip}} < \varepsilon^*\} \subset (m^* + m_0)\bar{\mathcal{C}}$$

and further $\text{Lip}_1(\mathbb{R}^d) \subset C_d \bar{\mathcal{C}}$ with $C_d := 2(m^* + m_0)/\varepsilon^*$. \square

Remark 3.6. Proposition 3.3 and Theorem 3.5 form the basis to prove (ii) of Theorem 2.3. It is worth pointing out that, with a suitable adaptation, the proof of the universal approximation theorem allows to show the density of $\bigcup_{m \geq 1} m\mathcal{C}$ in $\text{Lip}(\mathbb{R}^d)$ with respect to τ . However, we cannot prove something similar to Proposition 3.3 (ii) under the uniform norm. Therefore, Theorem 3.5, which proves that for the convex absorbing subset $\text{Lip}_1(\mathbb{R}^d)$ there exists m large enough such that $m\mathcal{C}$ is also dense in $\text{Lip}_1(\mathbb{R}^d)$, needs additional arguments.

Now we have all the ingredients to prove Theorem 2.3 (ii).

of Theorem 2.3 (ii). We assume first μ, ν have densities $u, v \in L^1(\mathbb{R}^d, (1 + |x|)dx)$ with respect to the Lebesgue measure. Recall Kantorovich's dual formulation (3.4), and one has

$$\begin{aligned} \mathcal{W}_1(\mu, \nu) &= \sup_{f \in \text{Lip}_1(\mathbb{R}^d)} \int_{\mathbb{R}^d} f(x)(u(x) - v(x)) dx \\ &\leq \sup_{f \in C_d \bar{\mathcal{C}}} \int_{\mathbb{R}^d} f(x)(u(x) - v(x)) dx \\ &= C_d \sup_{f \in \mathcal{C}} \int_{\mathbb{R}^d} f(x)(u(x) - v(x)) dx, \end{aligned}$$

where C_d is the constant in Theorem 3.5. As \mathcal{C} is the collection of convex combinations of functions $f(v \cdot x)$ with $v \in \mathbb{S}^{d-1}$ and $f \in \text{Lip}_1(\mathbb{R})$, it follows that

$$\begin{aligned} \mathcal{W}_1(\mu, \nu) &\leq C_d \sup_{f \in \mathcal{C}} \int_{\mathbb{R}^d} f(x)(u(x) - v(x)) dx \\ &= C_d \sup_{f \in \text{Lip}_1(\mathbb{R}), v \in \mathbb{S}^{d-1}} \int_{\mathbb{R}^d} f(v \cdot x)(u(x) - v(x)) dx \\ &= C_d \sup_{v \in \mathbb{S}^{d-1}} \mathcal{W}_1(\mu_v, \nu_v) = C_d \bar{\mathcal{W}}_1(\mu, \nu). \end{aligned}$$

Hence, (2.5) is established in view of Lemma 3.7. □

Lemma 3.7. *The subset of probability measures admitting a density is dense in $\mathcal{P}_1(\mathbb{R}^d)$ under \mathcal{W}_1 and $\bar{\mathcal{W}}_1$.*

Proof. Fix an arbitrary $\mu \in \mathcal{P}_1(\mathbb{R}^d)$ and take the density function $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}_+$ given by

$$\varphi(x) := \begin{cases} c \exp(1/(|x|^2 - 1)) & \text{if } |x| \leq 1 \\ 0 & \text{otherwise,} \end{cases}$$

where $c > 0$ is chosen such that $\int_{\mathbb{R}^d} \varphi(x) dx = 1$. Define the sequence of convolutions of measures $(\mu_t)_{t>0}$ by $\mu_t := \mu * \nu_t$, where the probability measure ν_t is identified by its density function $\varphi_t(x) := \varphi(x/t)/t^d$. By construction μ_t admits a density, and it remains to estimate $\mathcal{W}_1(\mu_t, \mu)$ according to (3.4). For each $f \in \text{Lip}_1(\mathbb{R}^d)$, it holds

$$\begin{aligned} \int_{\mathbb{R}^d} f(x)\mu_t(dx) - \int_{\mathbb{R}^d} f(x)\mu(dx) &:= \int_{\mathbb{R}^d \times \mathbb{R}^d} f(x+y)\mu(dx)\varphi_t(y)dy - \int_{\mathbb{R}^d} f(x)\mu(dx) \\ &= \int_{\mathbb{R}^d \times \mathbb{R}^d} (f(x+tz) - f(x))\mu(dx)\varphi(z)dz \\ &\leq t \int_{\mathbb{R}^d} z\varphi(z)dz \leq t, \end{aligned}$$

which implies $\mathcal{W}_1(\mu_t, \mu) \leq t$ and thus the desired density under \mathcal{W}_1 . The density under $\bar{\mathcal{W}}_1$ follows immediately from the inequality $\bar{\mathcal{W}}_1(\mu_t, \mu) \leq \mathcal{W}_1(\mu_t, \mu)$. □

3.3 Proof of Theorem 2.3 (iii)

To complete the proof of Theorem 2.3, we need an auxiliary result. Given a generic probability measure $\mu \in \mathcal{P}_p(\mathbb{R}^d)$, $\hat{\mu}^n$ is said to be its empirical measure (of order n) if

$$\hat{\mu}^n(dx) = \frac{1}{n} \sum_{k=1}^n \delta_{X^k}(dx),$$

where $(X^k)_{k \geq 1}$ is a sequence of i.i.d. random variables such that $X^k \sim \mu$ for all $k \geq 1$. By Theorem 1 of [8], there exist $C(p, d, M_p(\mu)) > 0$ and $\chi_{p,d} : \mathbb{N} \rightarrow \mathbb{R}_+$ such that

$$\mathbb{E}[\mathcal{W}_1(\mu, \hat{\mu}^n)] \leq C(p, d, M_p(\mu))\chi_{p,d}(n), \quad \text{for all } n \geq 1.$$

The function $\chi_{p,d}$ is specified in [8], while we need to refer to [10] for the explicit expression of $C(p, d, M_p(\mu))$. For $d = 1$ and $p = 3$, one has

$$\mathbb{E}[\mathcal{W}_1(\mu, \hat{\mu}^n)] \leq \frac{17280(M_3(\mu)^3 + 1)}{\sqrt{n}}, \quad \text{for all } n \geq 1. \tag{3.6}$$

of Theorem 2.3 (iii). We argue by contradiction. Assume that there exists $C > 0$ such that $\mathcal{W}_p(\mu, \nu) \leq C\mathcal{W}_p(\mu, \nu)$ for all $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$. Let $\ell^d \in \mathcal{P}_p(\mathbb{R}^d)$ be the Lebesgue measure concentrated on $[0, 1]^d$. Let $(G^n \equiv (G_1^n, \dots, G_d^n))_{n \geq 1}$ be a sequence of i.i.d. random variables distributed according to ℓ^d . Define $\hat{\mu}^n$ to be the empirical measure given by

$$\hat{\mu}^n(dx) := \frac{1}{n} \sum_{k=1}^n \delta_{G^k}(dx).$$

For any two sequences $(a_n)_{n \geq 1}, (b_n)_{n \geq 1} \subset \mathbb{R}_+$, we say $a_n \approx b_n$ if there exists a constant $c > 0$ such that

$$\frac{a_n}{c} \leq b_n \leq ca_n, \quad \text{for all } n \geq 1.$$

Then it follows by [1] that

$$\mathbb{E}[\mathcal{W}_1(\ell^d, \hat{\mu}^n)] \approx \begin{cases} (\log(n)/n)^{1/2} & \text{if } d = 2 \\ n^{-1/d} & \text{if } d \geq 3. \end{cases} \tag{3.7}$$

On the other hand, one has by assumption

$$\mathbb{E}[\mathcal{W}_1(\ell^d, \hat{\mu}^n)] \leq C\mathbb{E}[\mathcal{W}_1(\ell^d, \hat{\mu}^n)] = C \oint_{\mathbb{S}^{d-1}} \mathbb{E}[\mathcal{W}_1(\ell_v^d, \hat{\mu}_v^n)] dv,$$

where we recall that ℓ_v^d and $\hat{\mu}_v^n$ are the projections of ℓ^d and $\hat{\mu}^n$ along the direction $v = (v_1, \dots, v_d)$. Substituting ℓ_v^d and $\hat{\mu}_v^n$ into (3.6), one has

$$\mathbb{E}[\mathcal{W}_1(\ell_v^d, \hat{\mu}_v^n)] \leq \frac{17280(M_3(\ell_v^d)^3 + 1)}{\sqrt{n}} \leq \frac{17280(d^3 + 1)}{\sqrt{n}}, \quad \text{for all } v \in \mathbb{S}^{d-1},$$

where the second inequality follows from $M_3(\ell_v^d)^3 \leq M_3(\ell^d)^3 \leq d^3$. This yields

$$\mathbb{E}[\mathcal{W}_1(\ell^d, \hat{\mu}^n)] \leq \frac{17280C A_d (M_3(\ell_v^d)^3 + 1)}{\sqrt{n}},$$

and further, combined with (3.7), yields a contradiction for $d \geq 2$ and concludes the proof. \square

A

We start by recalling some elementary ingredients from functional analysis. Given a topological vector space \mathcal{E} , we denote by \mathcal{E}^* its dual space in separating duality via a bilinear form. The weak* convergence, denoted by w^* , is the convergence on \mathcal{E}^* induced by the elements of \mathcal{E} , i.e. $(e_n^*)_{n \geq 1} \subset \mathcal{E}^*$ converges to $e^* \in \mathcal{E}^*$ under w^* if and only if

$$\lim_{n \rightarrow \infty} \langle e_n^*, e \rangle = \langle e^*, e \rangle, \quad \text{for all } e \in \mathcal{E}.$$

Endowed with w^* , the dual space of \mathcal{E}^* is isometric to \mathcal{E} . In the following, we set $\mathcal{E} = L^1(\mathbb{R}^d)^{d+1}$ and $\mathcal{E}^* = L^\infty(\mathbb{R}^d)^{d+1}$ which are respectively the $(d + 1)$ -product of $L^1(\mathbb{R}^d)$ and $L^\infty(\mathbb{R}^d)$.

of Lemma 3.2. Note that $\text{Lip}(\mathbb{R}^d)$ embeds into the space $L^\infty(\mathbb{R}^d)^{d+1}$ via the map $\mathcal{L} : \text{Lip}(\mathbb{R}^d) \rightarrow L^\infty(\mathbb{R}^d)^{d+1}$ defined by

$$\mathcal{L}(f) := \left(\frac{f}{1 + |x|}, \nabla f := (\partial_1 f, \dots, \partial_d f) \right).$$

For each function $T : \text{Lip}(\mathbb{R}^d) \rightarrow \mathbb{R}$, define $T \circ \mathcal{L}^{-1} : \mathcal{L}(\text{Lip}(\mathbb{R}^d)) \rightarrow \mathbb{R}$ by

$$\mathcal{T}(\mathcal{L}(f)) := T(f).$$

Then by definition, T is τ -continuous linear if and only if $T \circ \mathcal{L}^{-1}$ is w^* -continuous linear. Following the arguments of Lemma 2.2 in [9], $\mathcal{L}(\text{Lip}(\mathbb{R}^d)) \subset L^\infty(\mathbb{R}^d)^{d+1}$, presented by weak equations

$$\mathcal{L}(\text{Lip}(\mathbb{R}^d)) = \left\{ (g, w) \in L^\infty(\mathbb{R}^d)^{d+1} : \partial_i((1 + |x|)g) = w_i, \partial_i w_j = \partial_j w_i, \text{ for } 1 \leq i, j \leq d \right\}$$

is a w^* -closed subspace. Then the dual spaces of $\mathcal{L}(\text{Lip}(\mathbb{R}^d))$ is included in the dual space of $L^1(\mathbb{R}^d)^{d+1}$, see e.g. page 129 in [4]. The proof is fulfilled by the fact that the elements (g, w) of $L^1(\mathbb{R}^d)^{d+1}$ represent all τ -continuous linear functions on $\mathcal{L}(\text{Lip}(\mathbb{R}^d))$ via

$$\mathcal{L}(\text{Lip}(\mathbb{R}^d)) \ni \mathcal{L}(f) \mapsto \int_{\mathbb{R}^d} g(x) \frac{f(x)}{1 + |x|} dx + \int_{\mathbb{R}^d} \nabla f(x) \cdot w(x) dx \in \mathbb{R}. \quad \square$$

References

[1] Luigi Ambrosio, Federico Glaudo, and Dario Trevisan, *On the optimal map in the 2-dimensional random matching problem*, Discrete & Continuous Dynamical Systems - A **39** (2019), 7291. MR-4026190

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou, *Wasserstein Gan*, arXiv:1701.07875, December 2017.

[3] N. Bonnotte, *Unidimensional and evolution methods for optimal transportation*, Ph.D. thesis, University of Paris-Sud, 2013.

[4] John B. Conway, *A course in functional analysis*, Graduate Texts in Mathematics, vol. 96, Springer-Verlag New York, 1985. MR-0768926

[5] G. Cybenko, *Approximations by superpositions of sigmoidal functions*, Mathematics of Control, Signals, and Systems **2** (1989), no. 4, 303–314. MR-1015670

[6] Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David A. Forsyth, and Alexander G. Schwing, *Max-sliced Wasserstein distance and its use for GANs*, Conference on Computer Vision and Pattern Recognition, 2019.

[7] L. Evans and R. Gariepy, *Measure theory and fine properties of functions*, Studies in Advanced Mathematics, CRC Press, 1991. MR-1158660

- [8] N. Fournier and A. Guillin, *On the rate of convergence in wasserstein distance of the empirical measure*, Probability Theory with Related Fields **162** (2015), 707–738. MR-3383341
- [9] G. Godefroy and N. Lerner, *Some natural subspaces and quotient spaces of l^1* , Advances in Operator Theory **3** (2018), no. 1, 61–74. MR-3730340
- [10] G. Guo and J. Obłój, *Computational methods for martingale optimal transport problems*, Annals of Applied Probability **29** (2019), no. 6, 3311–3347. MR-4047982
- [11] Juha Heinonen, *Lectures on analysis on metric spaces*, Universitext, Springer-Verlag, New York, 2001. MR-1800917
- [12] J. Johnson, *Banach space lipschitz functions and vector-valued lipschitz functions*, Bulletin of the American Mathematical Society **75** (1969), no. 6, 1334–1338. MR-0250048
- [13] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo K. Rohde, *Generalized sliced Wasserstein distances*, 33rd Conference on Neural Information Processing Systems, 2019.
- [14] F. Paty and M. Cuturi, *Subspace robust Wasserstein distances*, Proceedings of the 36th International Conference on Machine Learning, 2019.
- [15] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Berlot, *Wasserstein barycenter and its application to texture mixing*, Scale Space and Variational Methods in Computer Vision (Berlin, Heidelberg) (Alfred M. Bruckstein, Bart M. ter Haar Romeny, Alexander M. Bronstein, and Michael M. Bronstein, eds.), Springer Berlin Heidelberg, 2012, pp. 435–446.
- [16] Svetlozar T. Rachev and Ludger Rüschendorf, *Mass transportation problems volume 1: Theory*, Probability and Its Applications, Springer-Verlag New York, 1998. MR-1619171
- [17] Svetlozar T. Rachev and Ludger Rüschendorf, *Mass transportation problems volume 2: Applications*, Probability and Its Applications, Springer-Verlag New York, 1998. MR-1619171
- [18] Filippo Santambrogio, *Optimal transport for applied mathematicians*, Progress in Nonlinear Differential Equations and their Applications, vol. 87, Birkhäuser/Springer, Cham, 2015, Calculus of variations, PDEs, and modeling. MR-3409718
- [19] Cédric Villani, *Optimal transport: old and new*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 338, Springer-Verlag, Berlin, 2009. MR-2459454

Acknowledgments. We thank the anonymous referees for the thorough review and highly appreciate the comments and suggestions, which significantly contributed to improving the quality of the publication.

Electronic Journal of Probability

Electronic Communications in Probability

Advantages of publishing in EJP-ECP

- Very high standards
- Free for authors, free for readers
- Quick publication (no backlog)
- Secure publication (LOCKSS¹)
- Easy interface (EJMS²)

Economical model of EJP-ECP

- Non profit, sponsored by IMS³, BS⁴, ProjectEuclid⁵
- Purely electronic

Help keep the journal free and vigorous

- Donate to the IMS open access fund⁶ (click here to donate!)
- Submit your best articles to EJP-ECP
- Choose EJP-ECP over for-profit journals

¹LOCKSS: Lots of Copies Keep Stuff Safe <http://www.lockss.org/>

²EJMS: Electronic Journal Management System <http://www.vtex.lt/en/ejms.html>

³IMS: Institute of Mathematical Statistics <http://www.imstat.org/>

⁴BS: Bernoulli Society <http://www.bernoulli-society.org/>

⁵Project Euclid: <https://projecteuclid.org/>

⁶IMS Open Access Fund: <http://www.imstat.org/publications/open.htm>