

# HIGHER ORDER MDS CODES

S

ABSTRACT. We study the combinatorics of the Wasserstein-1 metric for various distances.

## 1. INTRODUCTION

The probability simplex

$$\Delta_{n-1} := \left\{ (p_1, \dots, p_n) \mid \sum_{i=1}^n p_i = 1 \text{ and } p_i \geq 0 \ \forall i = 1, \dots, n \right\}$$

consists of probability distributions of a discrete random variable with a state space of size  $n$ . We take this state space to be  $[n] := \{1, \dots, n\}$ . A *statistical model*  $\mathcal{M}$  is a subset of  $\Delta_{n-1}$  which represents distributions to which a hypothesized unknown distribution  $\boldsymbol{\nu}$  belongs. Typically, after collecting data  $\mathbf{u} = (u_1, \dots, u_n)$  where  $u_i$  is the number of times outcome  $i$  is observed, one forms the empirical distribution  $\bar{\boldsymbol{\mu}} = \frac{1}{N}\mathbf{u}$  where  $N = \sum_{i=1}^n u_i$  is the sample size. Note that  $\bar{\boldsymbol{\mu}} \in \Delta_{n-1}$ . To estimate the unknown distribution  $\boldsymbol{\nu}$ , a standard approach is to locate  $\boldsymbol{\nu} \in \mathcal{M}$ , that is a “closest” point to  $\bar{\boldsymbol{\mu}}$ . For instance,  $\boldsymbol{\nu}$  can be taken to be the maximum likelihood estimator [Sul18, Chapter 7] of  $\bar{\boldsymbol{\mu}}$ . In this case,  $\boldsymbol{\nu}$  is the point on  $\mathcal{M}$  that minimizes the Kullback-Leibler divergence from  $\bar{\boldsymbol{\mu}}$  to  $\mathcal{M}$ . However, Kullback-Leibler divergence is not a metric, and the maximum likelihood estimator does not minimize a true distance function from  $\bar{\boldsymbol{\mu}}$  to  $\mathcal{M}$ .

For the above density estimation problem, one can use a distance minimization approach if the state space  $[n]$  is also a metric space. A metric on  $[n]$  is a collection of nonnegative real numbers  $d_{ij}$  for  $i, j \in [n]$  such that  $d_{ii} = 0$  for all  $i \in [n]$ ,  $d_{ij} = d_{ji}$ , and the triangle inequality  $d_{ik} \leq d_{ij} + d_{jk}$  holds for all  $i, j, k \in [n]$ . Sometimes, the metric on  $[n]$  is written as an  $n \times n$  nonnegative symmetric matrix  $d = (d_{ij})_{i,j \in [n]}$ . Common examples include the discrete metric (all  $d_{ij} = 1$ ), the  $L_1$  metric ( $d_{ij} = |i - j|$ ), the  $L_0$  metric, and the Hamming distance metric.

For two probability distributions  $\boldsymbol{\mu}$  and  $\boldsymbol{\nu}$  in  $\Delta_{n-1}$ , the optimal value  $W_d(\boldsymbol{\mu}, \boldsymbol{\nu})$  of the following linear program is the *Wasserstein distance* between  $\boldsymbol{\mu}$  and  $\boldsymbol{\nu}$  based on the metric  $(d_{ij})$ :

$$(1) \quad \text{maximize} \quad \sum_{i=1}^n (\mu_i - \nu_i)x_i \quad \text{subject to} \quad |x_i - x_j| \leq d_{ij} \text{ for all } 1 \leq i < j \leq n.$$

This means we can define  $W_d(\boldsymbol{\mu}, \boldsymbol{\nu})$  for any pair of vectors  $\boldsymbol{\mu}, \boldsymbol{\nu}$  satisfying  $\mathbf{1}^\top \boldsymbol{\mu} = \mathbf{1}^\top \boldsymbol{\nu}$ . One should note that the constraint set of the variable  $\mathbf{x}$  in problem 1 is unbounded and that if  $\boldsymbol{\alpha} \in H_{n-1} := \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{1}^\top \mathbf{x} = 0\}$  and  $\lambda \in \mathbb{R}$  then  $\boldsymbol{\alpha}^\top (\mathbf{x} + \lambda \mathbf{1}) = \boldsymbol{\alpha}^\top \mathbf{x}$ . So we can equivalently formulate it as

$$W_d(\boldsymbol{\mu}, \boldsymbol{\nu}) = \max \{(\boldsymbol{\mu} - \boldsymbol{\nu})^\top \mathbf{x} \mid \mathbf{x} \in H_{n-1}, |x_i - x_j| \leq d_{ij} \forall i, j\}$$

which has a bounded constraint set. The constraint set of this linear program is called the *Lipshitz polytope*

$$P_d = \{\mathbf{x} \in H_{n-1} \mid |x_i - x_j| \leq d_{ij} \forall 1 \leq i < j \leq n\}.$$

The Wasserstein distance  $W_d(\boldsymbol{\mu}, \mathcal{M})$  from  $\boldsymbol{\mu} \in \Delta_{n-1}$  to a set  $\mathcal{M}$  is the infimum of  $W_d(\boldsymbol{\mu}, \boldsymbol{\nu})$  as  $\boldsymbol{\nu}$  ranges over  $\mathcal{M}$ :

$$(2) \quad W_d(\boldsymbol{\mu}, \mathcal{M}) := \min_{\boldsymbol{\nu} \in \mathcal{M}} W_d(\boldsymbol{\mu}, \boldsymbol{\nu}).$$

This has been successfully used to construct a version of Generative Adversarial Networks [ACB17] where  $W_d(\cdot, \mathcal{M})$  is used as the loss function. However, for large  $n$ , computing  $W_d(\boldsymbol{\mu}, \mathcal{M})$  exactly is not feasible with the current state of knowledge. If we take  $\mathcal{M} = \{\boldsymbol{\nu}\}$  we recover the original Wasserstein distance  $W_d(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min \{\lambda \geq 0 \mid \boldsymbol{\nu} \in \boldsymbol{\mu} + \lambda B\}$ .

In this paper our starting point is [ÇJM+20; ÇJM+21] to study the combinatorics of the Wasserstein unit ball. Such combinatorics governs the combinatorial complexity (contrast against algebraic complexity) of problem 2. We first recall this approach.

The Wasserstein distance  $W_d$  induced by the finite metric  $d$  on  $[n]$  defines a norm on  $H_{n-1}$  namely

$$\|\boldsymbol{\alpha}\|_d = \|\boldsymbol{\alpha}\|_d^W = \max \{\boldsymbol{\alpha}^\top \boldsymbol{\mu} \mid \mathbf{x} \in H_{n-1}, |x_i - x_j| \leq d_{ij} \forall 1 \leq i < j \leq n\}.$$

The unit ball of this norm is the polytope

$$(3) \quad B_d = \text{conv} \left\{ \frac{1}{d_{ij}} (\mathbf{e}_i - \mathbf{e}_j) : 1 \leq i < j \leq n \right\},$$

where  $B$  lies in the hyperplane  $H_{n-1}$  and is the dual of the *Lipshitz polytope*  $P_d$ . It is well known that the  $k$  dimensional facets of  $P_d$  are in on-to-one correspondence with the  $k$  codimensional facets of  $B_d$ . In other words, the number of  $k$  dimensional facets of  $P_d$  is equal to the number of  $n - 2 - k$  dimensional facets of  $B_d$ .

## 2. VERTICES OF $B_d$ WITH $d$ INDUCED BY A GRAPH

Consider the discrete metric  $d$  on  $[n]$ . Formally this is given by  $d_{ij} = 1 \forall i \neq j$ . [CM14; ÇJM+21] prove that the number of  $k$  dimensional facets of  $B_d$  is  $\binom{n}{k+2} (2^{k+2} - 2)$ . In particular, the number of vertices ( $k = 0$ ) is  $n(n-1)$ . This is the maximum number of possible vertices a Wasserstein ball can have, for any metric  $d$ , by the description in Equation (3). Here is an alternate way to think about the metric  $d$ . Consider the complete graph  $K_n$  on  $n$  vertices, labelled with  $[n]$ , so every vertex is connected to every other vertex

by an edge. Then  $d_{ij} = 1$  is the length of the shortest path to reach  $j$  from  $i$  on this graph. This graph has precisely  $\binom{n}{2}$  edges. Soon it will turn out that the number of vertices of  $B_d$  being double the number of edges is not a coincidence. Further, based on this example, we propose the following definition.

**Definition 2.1** (Wasserstein metric based on a graph). *Let  $G = ([n], E, w)$  be a connected weighted undirected graph without self loops that has vertices  $[n]$ , edges  $E$  and non-negative weights given by  $w : E \rightarrow \mathbb{R}_{\geq 0}$ . If  $G$  is unweighted, we simply treat  $G$  as a weighted graph with weights of all edges as 1. Define  $d_{ij}$  to be the weighted length of the shortest path from vertex  $i$  to  $j$ . The Wasserstein metric  $W_G$  based on graph  $G$  is defined to be the Wasserstein metric  $W_d$  based on  $d$ .*

Corresponding to the abovementioned Wasserstein metric  $W_G$ , its unit ball in  $H_{n-1}$  will be denoted by  $B_G$ .

*Example 2.2.* The metric induced by an unweighted line graph on  $n$  vertices is said to be the  $L_1$  metric on  $[n]$ . Let's look at  $n = 3$ . So  $G$  is  $1-2-3$ . The corresponding metric is given by  $d_{ij} = |i - j|$ . According to Equation (3),  $B_G$  is the convex hull of the points  $\mathbf{u}_{\pm} = \pm(1, -1, 0)$ ,  $\mathbf{v}_{\pm} = \pm(0, 1, -1)$ ,  $\mathbf{w}_{\pm} = \pm(0.5, 0, -0.5)$ . But  $\mathbf{w}_{\pm} = \frac{1}{2}\mathbf{u}_{\pm} + \frac{1}{2}\mathbf{v}_{\pm}$  hence not vertices. The vertices of  $B_G$  turn out to be exactly  $\mathbf{u}_{\pm}, \mathbf{v}_{\pm}$ ; so total 4 in number. Again observe that the number of vertices of  $B_G$  is double the number of edges in  $G$ .

Next we will turn towards the key result in this section, namely the phenomenon we observed both for the discrete and  $L_1$  metric.

**Theorem 2.3.** *Let  $G = ([n], E)$  be a connect unweighted undirected graph without self loops on  $n$  vertices. Then the unit ball  $B_G$  of the Wasserstein metric induced by  $G$  has precisely  $|E|$  vertices.*

Before starting the proof right away, we present an observation that was key in the examples of discrete and  $L_1$  metrics. Our graph  $G$  is connected, unweighted and undirected. If shortest path from  $i$  to  $j$  is  $i = x_1 \rightarrow x_2 \rightarrow \dots \rightarrow x_p = j$  then  $d_{ij} = p - 1$  and  $\frac{\mathbf{e}_j - \mathbf{e}_i}{d_{ij}} = \frac{\mathbf{e}_j - \mathbf{e}_i}{p - 1} =$

$\frac{1}{p - 1} \sum_{t=1}^{p-1} (\mathbf{e}_{x_{t+1}} - \mathbf{e}_{x_t}) = \frac{1}{p - 1} \sum_{t=1}^{p-1} \frac{\mathbf{e}_{x_{t+1}} - \mathbf{e}_{x_t}}{d_{x_t x_{t+1}}}$ . In other words,  $\frac{\mathbf{e}_j - \mathbf{e}_i}{d_{ij}}$  is never a vertex of  $B_G$  because it is a convex combination of some other points in  $B_G$  corresponding to edges in  $G$ .

If we want to determine a  $d$ , for given  $n$  and number of vertices  $2\alpha$ , for which the constraint matrix  $M$  satisfies that its rank is  $2\alpha$ , we want to find a rank 2 matrix  $M$  with the rows being  $\frac{\mathbf{e}_i - \mathbf{e}_j}{d_{ij}}$ , such that its rank is  $2\alpha$ , then equivalently we want to search for a matrix  $X = M^T M \succeq 0$  with rank  $2\alpha$ .

## REFERENCES

- [CM14] P. Cellini and M. Marietti. “Root polytopes and Abelian ideals”. In: *Journal of Algebraic Combinatorics* 39.3 (2014), pp. 607–645. DOI: [10.1007/s10801-013-0458-5](https://doi.org/10.1007/s10801-013-0458-5). URL: <https://doi.org/10.1007/s10801-013-0458-5>.
- [ACB17] M. Arjovsky, S. Chintala, and L. Bottou. “Wasserstein Generative Adversarial Networks”. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, June 2017, pp. 214–223. URL: <https://proceedings.mlr.press/v70/arjovsky17a.html>.
- [Sul18] S. Sullivant. *Algebraic statistics*. Vol. 194. American Mathematical Soc., 2018.
- [ÇJM+21] T. Ö. Çelik, A. Jamneshan, G. Montúfar, B. Sturmfels, and L. Venturello. “Wasserstein distance to independence models”. In: *Journal of symbolic computation* 104 (2021), pp. 855–873.
- [ÇJM+20] T. Ö. Çelik, A. Jamneshan, G. Montúfar, B. Sturmfels, and L. Venturello. “Optimal transport to a variety”. In: *Mathematical aspects of computer and information sciences*. Vol. 11989. Lecture Notes in Comput. Sci. Springer, Cham, [2020] ©2020, pp. 364–381.