CHARACTERIZING VERTICES OF WAASSERSTEIN BALL

S

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

1. Introduction

The probability simplex

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

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 Δ_{n-1} which represents distributions to which a hypothesized unknown distribution $\boldsymbol{\nu}$ belongs. Typically, after collecting data $\boldsymbol{u} = (u_1, \cdots, u_n)$ where u_i is the number of times outcome i is observed, one forms the empirical distribution $\bar{\boldsymbol{\mu}} = \frac{1}{N}\boldsymbol{u}$ 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 Δ_{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}):

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

Date: July 2024.

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 $\boldsymbol{x}$ in problem 1 is unbounded and that if $\boldsymbol{\alpha} \in H_{n-1} := \left\{\boldsymbol{x} \in \mathbb{R}^n \;\middle|\; \mathbf{1}^\top\boldsymbol{x} = 0\right\}$ and $\lambda \in \mathbb{R}$ then $\boldsymbol{\alpha}^\top(\boldsymbol{x}+\lambda\mathbf{1}) = \boldsymbol{\alpha}^\top\boldsymbol{x}$. So we can equivalently formulate it as

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

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

$$P_d = \left\{\boldsymbol{x} \in H_{n-1} \;\middle|\; |x_i - x_j| \leq d_{ij}\ \forall\ 1 \leq i < j \leq n\right\}.$$

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

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

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 $W_d(\boldsymbol{\mu},\boldsymbol{\nu}) = \min\left\{\lambda \geq 0 \;\middle|\; \boldsymbol{\nu} \in \boldsymbol{\mu} + \lambda B\right\}$.

In this paper our starting point is [ÇJM+20; ÇJM+21] to study the combinatorics of the Wasserstein unit ball. Such combinatorics is 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

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

The unit ball of this norm is the polytope

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

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 $w : E^2 \to \mathbb{R}_{\geq 0}$. 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 $\boldsymbol{u}_{\pm} = \pm(1, -1, 0), \boldsymbol{v}_{\pm} = \pm(0, 1, -1), \boldsymbol{w}_{\pm} = \pm(0.5, 0, -0.5)$. But $\boldsymbol{w}_{\pm} = \frac{1}{2}\boldsymbol{u}_{\pm} + \frac{1}{2}\boldsymbol{v}_{\pm}$ hence not vertices. The vertices of B_G turn out to be exactly $\boldsymbol{u}_{\pm}, \boldsymbol{v}_{\pm}$; 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. Such results have been studied for weighted graphs in [MP22, Theorem 2, §3.1], however our proof technique is purely combinatorial and constructions are slightly different.

Theorem 2.3. *Let $G = ([n], E)$ be a connected unweighted undirected graph without self loops on n vertices. Then the unit ball B_G of the Wasserstein metric induced by G has precisely $2|E|$ vertices, namely $\{\boldsymbol{e}_i - \boldsymbol{e}_j \mid \{i, j\} \in E\}$.*

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 $i = x_1 \to x_2 \to \cdots \to x_p = j$ then $d_{ij} = p - 1$ and $\dfrac{\boldsymbol{e}_j - \boldsymbol{e}_i}{d_{ij}} = \dfrac{\boldsymbol{e}_j - \boldsymbol{e}_i}{p-1} = \dfrac{1}{p-1}\sum_{t=1}^{p-1}(\boldsymbol{e}_{t+1} - \boldsymbol{e}_t) = \dfrac{1}{p-1}\sum_{t=1}^{p-1}\dfrac{\boldsymbol{e}_{x_{t+1}} - \boldsymbol{e}_{x_t}}{d_{x_t x_{t+1}}}$. In other words, $\dfrac{\boldsymbol{e}_j - \boldsymbol{e}_i}{d_{ij}}$ 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 2α, for which the constraint matrix M satisfies that its rank is 2α, we want to find a rank 2 matrix M with the rows being $\frac{\boldsymbol{e}_i - \boldsymbol{e}_j}{d_{ij}}$, such that its rank is 2α, then equivalently we want to search for a matrix $X = M^\top M \succeq 0$ with rank 2α.

Proof. By the above observation, if $\{i,j\} \notin E$ then $\dfrac{\boldsymbol{e}_i - \boldsymbol{e}_j}{d_{ij}}$ is not a vertex. We now need to verify that $\dfrac{\boldsymbol{e}_i - \boldsymbol{e}_j}{d_{ij}}$, for $\{i,j\} \in E$, are indeed vertices. Suppose otherwise that, WLOG $\{1,2\} \in E$ but $\boldsymbol{e}_1 - \boldsymbol{e}_2$ is not a vertex. Then $\exists$ edges $\{s_1,t_1\}, \cdots, \{s_k,t_k\} \in E \setminus \{\{1,2\}\}$ and $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_k) \in \Delta_{k-1}$ with all positive entries such that $\boldsymbol{e}_1 - \boldsymbol{e}_2 = \sum_{i=1}^{k} \lambda_i(\boldsymbol{e}_{s_i} - \boldsymbol{e}_{t_i})$. Let $S_j \coloneqq \{i \in [k] \mid s_i = j\}$ and $T_j \coloneqq \{i \in [k] \mid t_i = j\}$ for every $j \in [n]$. Note that $S_j \cap T_j = \varnothing$, otherwise if $r \in S_j \cap T_j$ then $s_r = t_r = j$ but we had assumed G has no self loops. Multiplying $\boldsymbol{e}_1^\top$ on both sides gives $1 = \displaystyle\sum_{i \in S_1} \lambda_i - \sum_{i \in T_1} \lambda_i$. We have thus written 1 as a difference of two non-negative numbers both of which are ≤ 1. This forces $\sum_{i \in S_1} \lambda_i = 1, \sum_{i \in T_1} \lambda_i = 0$. The former consequence means that $S_1 = [k]$ because we had assumed each $\lambda_i > 0$. In other words $s_i = 1 \ \forall\ i \in [k]$. The same reasoning when the equation is multiplied with $\boldsymbol{e}_2^\top$ gives $T_2 = [k]$. So $t_i = 2 \ \forall\ i \in [k]$. This just reduces each edge $\{s_i, t_i\}$ to be $\{1,2\}$ which is false by hypothesis. Therefore $\boldsymbol{e}_1 - \boldsymbol{e}_2$ must be a vertex of B_G if $\{1,2\} \in E$. The same argument applies for any $\boldsymbol{e}_i - \boldsymbol{e}_j$ if $\{i,j\} \in E$. $\blacksquare$

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. 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.

[MP22] L. Montrucchio and G. Pistone. "Kantorovich distance on finite metric spaces: Arens–Eells norm and CUT norms". In: *Information Geometry* 5.1 (2022), pp. 209–245. DOI: 10.1007/s41884-021-00050-w. URL: https://doi.org/10.1007/s41884-021-00050-w.

[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.