A Prelude on the Wasserstein Metric

Ryan Tay

14 October 2024

Suppose we have probability measures μ and ν on a sample space X. How should one quantify how "different" the two probability measures are? If the sample space is $X = \mathbb{R}$ and the measures μ and ν admit probability density functions f and g respectively, so that $\mu(A) = \int_A f(x)\,\mathrm{d}x$ and $\nu(A) = \int_A g(x)\,\mathrm{d}x$, then a quantity which could measure the "distance" between μ and ν is

$$\|f - g\|_{L^1} \coloneqq \int_{\mathbb{R}} |f(x) - g(x)|\,\mathrm{d}x.$$

If $\|f - g\|_{L^1}$ is close to 0 then we would say that the two measures are really close to each other, and if $\|f - g\|_{L^1}$ is close to 2 then we would say that the two measures are really quite different from each other.

There are two problems with this approach. Firstly, some measures may not even admit a probability density function. For instance, the Dirac measure δ_0 on $\mathbb{R}$ given by

$$\delta_0(A) \coloneqq \begin{cases} 1 & \text{if } 0 \in A, \\ 0 & \text{else.} \end{cases}$$

The elephant in the room, though, is that the L^1 distance between f and g may not match up with our intuitive notion of how far measures μ and ν may be from each other. Consider, for instance, the measures μ and ν having densities

$$f(x) \coloneqq \begin{cases} 10^8 & \text{if } 0 \le x \le \frac{1}{10^8}, \\ 0 & \text{else,} \end{cases} \qquad \text{and} \qquad g(x) \coloneqq \begin{cases} 10^8 & \text{if } \frac{1}{10^8} \le x \le \frac{2}{10^8}, \\ 0 & \text{else,} \end{cases}$$

respectively. Then $\|f - g\|_{L^1} = 2$, which as large as $\|f - g\|_{L^1}$ could possibly be, and so we would say that μ and ν are "really far apart from each other". There is, however, still a sense in which μ and ν are really close to each other: the graph of g is just the graph of f translated to the right by a mere 10^{-8} units. Furthermore, if we consider a third measure ξ with density

$$h(x) \coloneqq \begin{cases} 10^8 & \text{if } 500 \le x \le 500 + \frac{1}{10^8}, \\ 0 & \text{else,} \end{cases}$$

then we would also have $\|f - h\|_{L^1} = 2$. The L^1 distance is unable to capture this huge shift of 500 units to the right; it is unable to distinguish μ from ξ any more than it can distinguish μ from ν.

Instead, we turn to a way of quantifying the distance between two probability measures μ and ν based off of this intuitive notion of how much "effort" would it take to "move" from μ to ν. Underpinning this whole theory is the Monge-Kantorovich problem [Gar18, Chapter 20.2] [RR, Chapter 2.1], named after Gaspard Monge and Leonid Kantorovich (Russian: Леонид Канторович):

The Monge-Kantorovich problem.
Fix Polish spaces X and Y, and a lower semicontinuous cost function $c\colon X \times Y \to$

$\mathbb{R}_{\geq 0}$. Given two Borel probability measures μ and ν on X and Y respectively, the Monge-Kantorovich problem seeks to minimise[1] the following total cost:

$$\int_{X \times Y} c \, \mathrm{d}\pi,$$

with π ranging over the space of all Borel probability measures on $X \times Y$ with marginals μ and ν.

An intuitive view of the Monge-Kantorovich problem is as follows. Suppose you had a pile of sand spread about in a space X according to a probability measure μ, and you wished to transport that pile of sand to a space Y and spreading it out according to a probability measure ν. Moving one particle of sand from $x \in X$ to $y \in Y$ costs $\$c(x, y)$. A transport plan π tells you how you should move the pile of sand: a volume of space $B \subseteq Y$ gets $\pi(A \times B)$ of the sand available the volume of space $A \subseteq X$. Of course, there are restrictions for what constitutes a transport plan: after the transportation has finished, the amount of sand in any $B \subseteq Y$ should equal $\pi(X \times B)$. Similarly, the amount of sand leaving any $A \subseteq X$ should equal $\pi(A \times Y)$.

It is these restrictions which are captured by the requirement that π has marginals μ and ν, that is, given the projection maps $p_X \colon X \times Y \to X$ and $p_Y \colon X \times Y \to Y$ defined by

$$p_X(x, y) \coloneqq x \quad \text{and} \quad p_Y(x, y) \coloneqq y,$$

we require that μ is the image measure of π under the map p_X, and ν is the image measure of π under the map p_Y.

A transport plan always exists: the product measure $\pi \coloneqq \mu \times \nu$ works. The Monge-Kantorovich problem, however, seeks to find the optimal transport plan. It is an important result that the Monge-Kantorovich problem *always* has a solution, in the sense that an optimal transport plan always exists (see **??**). It is this fact that allows us to develop the Wasserstein metric[2].

For a topological space X, let us use $P(X)$ to denote the space of Borel probability measures on X, though we will mainly be concerned with the spaces $P(\mathbb{R})$ and $P(\mathbb{R}^2)$. For $\mu, \nu \in P(\mathbb{R})$, we define the space

$$\Pi_{\mu,\nu} \coloneqq \{ \, \pi \in P(\mathbb{R}^2) : \pi \text{ has marginals } \mu \text{ and } \nu \, \}.$$

We shall use the cost function $c \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ given by $c(x, y) \coloneqq |x - y|$ to define the Wasserstein metric, named after Leonid Vaserstein[3] (Russian: Леонид Васерштейн), on the subspace

$$P_1(\mathbb{R}) \coloneqq \left\{ \mu \in P(\mathbb{R}) : \int_{\mathbb{R}} |x| \, \mathrm{d}\mu(x) < \infty \right\}.$$

Definition 1 (The Wasserstein metric W_1 [Gar18, Chapter 21] [Vil09, Definition 6.1, Definition 6.4])

[1] A priori, we seek to find the infimum value of $\int_{X \times Y} c(x, y) \, \mathrm{d}\pi$, rather than the minimum value. It will turn out that the infimum value is actually always achieved (see **??**).

[2] Also known as the Kantorovich-Rubinstein metric or the Earth Mover's distance [Vil09, Chapter 6 Bibliographical Notes].

[3] Leonid Vaserstein was actually not the first person to come up with the Wasserstein metric. The metric is due to Leonid Kantorovich and Gennadii Rubinstein (Russian: Геннадии Рубинштейн) [Vil09, Chapter 6 Bibliographical Notes] [Vil09, Chapter 3].

We define the function $W_1 \colon P_1(\mathbb{R}) \times P_1(\mathbb{R}) \to \mathbb{R}_{\geq 0}$*, called the Wasserstein metric*[4]*, by*

$$W_1(\mu, \nu) \coloneqq \inf \left\{ \int_{\mathbb{R}^2} |x - y| \, \mathrm{d}\pi(x, y) : \pi \in \Pi_{\mu,\nu} \right\}.$$

The Wasserstein metric W_1 is indeed a metric on $P_1(\mathbb{R})$ (see **??**), so it captures the ideas of "distance" between measures as one would expect, including the triangle inequality.

Calculating $W_1(\mu, \nu)$ for specific probability measures μ and ν can be challenging given only this definition. Often, if there is enough similarity between μ and ν, we can obtain an upper bound for $W_1(\mu, \nu)$ by finding a suitable transport plan from μ to ν. Recall again the measures from the Prelude section:

$$\mu(A) \coloneqq 10^8 \cdot \lambda \left(A \cap \left[0, \frac{1}{10^8} \right] \right) \quad \text{and} \quad \nu(A) \coloneqq 10^8 \cdot \lambda \left(A \cap \left[\frac{1}{10^8}, \frac{2}{10^8} \right] \right),$$

where λ is the Lebesgue measure on $\mathbb{R}$. Consider the transport plan $T \colon \mathbb{R} \to \mathbb{R}$ given by $T(x) \coloneqq x + \frac{1}{10^8}$. Let $G \colon \mathbb{R} \to \mathbb{R}^2$ be defined by $G(x) \coloneqq (x, T(x))$, and define the Borel probability measure π on $\mathbb{R}^2$ to be the image measure of μ under G, that is,

$$\pi(A) \coloneqq \mu(G^{-1}(A)).$$

Then π has marginals μ and ν, and we observe that

$$\int_{\mathbb{R}^2} |x - y| \, \mathrm{d}\pi(x, y) = \frac{1}{10^8},$$

yielding $0 < W_1(\mu, \nu) \leq \frac{1}{10^8}$. Given that the graph of the distribution of ν is just a translation of $\frac{1}{10^8}$ units to the right of the graph of the distribution of μ, would it not be nice if $W_1(\mu, \nu) = \frac{1}{10^8}$? Furthermore, if we recall the other Borel probability measure ξ from the Prelude section, namely

$$\xi(A) \coloneqq 10^8 \cdot \lambda \left(A \cap \left[500, \ 500 + \frac{1}{10^8} \right] \right),$$

a similar argument to the one above would yield $0 < W_1(\mu, \xi) \leq 500$. Again, it would be wonderful if we indeed had $W_1(\mu, \xi) = 500$.

Theorem 2 ([Gar18, Corollary 21.2.3])
If two Borel probability measures μ *and* ν *on* $\mathbb{R}$ *have respective probability density functions* f_μ *and* f_ν*, then*

$$W_1(\mu, \nu) \geq \left| \int_{\mathbb{R}} x f_\mu(x) \, \mathrm{d}x - \int_{\mathbb{R}} x f_\nu(x) \, \mathrm{d}x \right|.$$

In other words, the Wasserstein distance between μ *and* ν *is at least the distance between their means.*

Recall our example Borel probability measures on $\mathbb{R}$:

$$\begin{aligned}
\mu(A) &\coloneqq 10^8 \cdot \lambda \left(A \cap \left[0, \frac{1}{10^8} \right] \right) \\
\nu(A) &\coloneqq 10^8 \cdot \lambda \left(A \cap \left[\frac{1}{10^8}, \frac{2}{10^8} \right] \right), \quad \text{and} \\
\xi(A) &\coloneqq 10^8 \cdot \lambda \left(A \cap \left[500, \ 500 + \frac{1}{10^8} \right] \right).
\end{aligned}$$

Theorem 2 together with our earlier discussions yield $W_1(\mu, \nu) = \frac{1}{10^8}$ and $W_1(\mu, \xi) = 500$.

[4] The appearance of the subscript "1" in the notations "W_1" and "P_1" is due to the definition of the more general Wasserstein p-metric W_p, defined by

$$W_p(\mu, \nu) \coloneqq \left(\inf \left\{ \int_{\mathbb{R}^2} |x - y|^p \, \mathrm{d}\pi(x, y) : \pi \in \Pi_{\mu,\nu} \right\} \right)^{1/p},$$

where $1 \leq p < \infty$. This metric W_p will be defined on the space $P_p(\mathbb{R})$ consisting of all $\mu \in P(\mathbb{R})$ such that $\int_{\mathbb{R}} |x|^p \, \mathrm{d}\mu(x) < \infty$.

Bibliography and References

[Aki22] Akira. Gluing lemma in optimal transport. Mathematics Stack Exchange. URL https://math.stackexchange.com/q/4489709, 2022.

[Eva24] Josephine Evans. Personal communication, 2024.

[Gar18] David Garling. Analysis on Polish Spaces and an Introduction to Optimal Transportation. Cambridge University Press, 2018.

[Led01] Michel Ledoux. The Concentration of Measure Phenomenon. Volume 89 of Mathematical Surveys and Monographs. American Mathematical Society, 2001.

[MN11] Flavia-Corina Mitroi and Constantin Niculescu. An extension of Young's inequality. Abstract and Applied Analysis, 2011, 2011.

[RR] Svetlozar Rachev and Ludger Rüdenschorf. Mass Transportation Problems. Probability and its Applications. Springer New York, NY.

[San15] Filippo Santambrogio. Optimal Transport for Applied Mathematicians. Progress in Nonlinear Differential Equations and Their Applications. Birkhäuser Cham, 2015.

[Vil03] Cédric Villani. Topics in Optimal Transportation. Volume 58 of Graduate Studies in Mathematics. American Mathematical Society, 2003.

[Vil09] Cédric Villani. Optimal Transport: Old and New. Springer-Verlag Berlin Heidelberg, 2009.