An Overview of Optimal Transport Theory and the Wasserstein Distance

Austin Tao

Columbia University
Department of Applied Physics and Applied Mathematics

March 23, 2023

1. Introduction

A distance function $\mathbf{d}(a, b)$ can be thought of as a bivariate operator that maps to the set of positive real numbers. This distance function is also a metric if it satisfies the following properties:

1. $\mathbf{d}(a, b) \geq 0$
2. $\mathbf{d}(a, b) = 0$ if and only if $a = b$
3. $\mathbf{d}(a, b) = \mathbf{d}(b, a)$
4. $\mathbf{d}(a, b) \leq \mathbf{d}(a, c) + \mathbf{d}(b, c)$

The most common example of such a distance function is the L_2 distance, otherwise known as the Euclidean distance, which is defined as follows:

$$\mathbf{d}_2(a, b) = ||a - b|| = \sqrt{\sum_{i=1}^{d} (a_i - b_i)^2} \tag{1}$$

Other distance metrics exist, including the Hellinger distance, L_∞ distance, and more. However, there are some drawbacks to using these distance metrics:

1. Usually, we cannot compare two sets P and Q when one set is discrete and the other is continuous.
2. Some distance metrics are highly sensitive to tiny changes in the distribution(s)
3. When averaging similar objects (say, a distribution), these metrics don't always yield a similar object.

This is where the Wasserstein Distance becomes useful – none of these drawbacks are present when using the Wasserstein Distance metric.

2. The Wasserstein Distance

2.1 Optimal Transport Theory

To understand the intuition behind the Wasserstein Distance, it is necessary to first consider Optimal Transport Theory, first studied by Garspard Monge in the 18th century. He wanted to determine what the best strategy was for transforming one pile of sand into another pile of sand (not necessarily with the same structure). In other words, what is the transportation plan that minimizes some cost function? This problem is expressed mathematically as follows:

$$\inf_{T:X\to X} \int_X c(x, T(x))u(x)dx \tag{2}$$

subject to the following constraint:

$$\forall B \subset X, \int_{T^{-1}(B)} u(x)dx = \int_B v(x)dx \tag{3}$$

Here, X is a subset of $\mathbb{R}^2$, the positive functions u and v represent the piles of sand, and $c(x, T(x))$ represents the cost function. The problem is now understood as finding a transport map $T(x)$ that transforms u into v (while obeying conservation of mass) while minimizing the product of $c(x, T(x))$ and $u(x)$. When written with measures, equations (2) and (3) can be written as follows:

$$\inf_{T:X\to Y} \int_X c(x, T(x))d\mu \tag{4}$$

subject to

$$v = T\#\mu \tag{5}$$

Here, X and Y are measurable sets, we have $\mu(X) = v(Y)$, and the constraint that T pushes μ onto v (conservation of mass constraint). If a minimizer $T : X \to Y$ exists, then that is the optimal transport map. However, this minimizer might not exist, which leads us to the Kantorovich formulation where mass at location x is allowed to move to more than one location. For instance, a point with weight 0.2 could split into two points each with weight 0.1 and move in different directions.

2.2 Defining the Wasserstein Distance

Now, let $\Gamma(P, Q)$ denote all joint distributions J for (X, Y) that have marginals P and Q. Then, the Wasserstein Distance is

$$W_p(P, Q) = \left(\inf_{J\in\Gamma(P,Q)} \int ||x - y||^p dJ(x, y)\right)^{1/p} \tag{6}$$

where $p \geq 1$, and the minimizer J is the optimal transport plan. This generalizes for any distance metric $d(x, y)$ as follows:

$$W_p(P, Q) = \left(\inf_{J\in\Gamma(P,Q)} \int d(x, y)^p dJ(x, y)\right)^{1/p} \tag{7}$$

Looking back on the properties a distance function must satisfy to be a metric, we see that the Wasserstein Distance is indeed a metric:

1. The range of the Wasserstein Distance is $[0, \infty]$, meaning that we have $W(a, b) \geq 0$ in all cases.
2. It is true that $W(a, b) = 0$ if and only if $a = b$; this can be verified by considering two distributions a and b. If $a = b$, this means the two distributions are one and the same, thus resulting in the Wasserstein Distance between them to be zero. If $W(a, b) = 0$, we know that the distributions cannot be different by any amount, meaning they must be one and the same.
3. The symmetry of the Wasserstein distance is obvious. Transforming one distribution into the second is identical to transforming the second distribution into the first.
4. The proof that the Wasserstein distance satisfies the Triangle Inequality can be found by referring to [2], which gives an elementary proof on the matter.

2.3 Example Cases

Example 1: Consider the case where we have only point masses (degenerate distributions). Let $u = \delta_{a_1}$ and $v = \delta_{a_2}$ both be degenerate distributions (meaning a distribution equal to zero everywhere except for zero and whose integral over the real line is one) located at points a_1 and a_2 in $\mathbb{R}$. Since they are degenerate distributions, there can only be one possible coupling of these two measures. This is located at (a_1, a_2) and is a point mass. If we then use the absolute value function as the distance function in our Wasserstein distance calculation, we find that for any $p \geq 1$, the p-th Wasserstein Distance between the two point masses is given by

$$W_p(u, v) = |a_1 - a_2| \tag{8}$$

Example 2: Now, consider the case where we have two normal distributions on $\mathbb{R}^n$ i.e. $P = N(\mu_1, \Sigma_1)$ and $Q = N(\mu_2, \Sigma_2)$. Using the Euclidean norm as the distance function, we can then write the 2-Wasserstein distance between P and Q as follows:

$$W_2(P, Q)^2 = |\mu_1 - \mu_2|^2 + \text{tr}(\Sigma_1) + \text{tr}(\Sigma_2) - 2\text{tr}\left[(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right] \tag{9}$$

where $\text{tr}(A)$ represents the trace of the matrix A. Notice that this solution turns into equation (8) in the case where $p = 2$ and we only have two point masses because the trace terms drop out, leaving only $|\mu_1 - \mu_2|^2$.

Unfortunately, the one dimensional case and the case of two Gaussian measures are the only simple cases for actually calculating the Wasserstein Distance. Computing the W_2 distance in higher dimensions requires computational numerical methods, which is outside the scope of this paper.

2.4 Properties of the Wasserstein Distance

Considering the W_2 distance, one important property is that the squared W_2 metric is jointly convex in translations and dilations of the input data. Specifically, this is described in [3]:

Let f and g be compactly supported probability density functions with finite second order moment on an interval $\Omega \subset \mathbb{R}$. Then,

$$W_2^2(f(t-s), g(t)) = W_2^2(g(t), f(t)) + s^2 + 2s \int_\Omega (x - T(x)) f(x) dx \tag{10}$$

where $s \in \mathbb{R}$ and T is the optimal map from f to g. Furthermore,

$$W_2^2(f(t), Af(At - s)) \tag{11}$$

is convex in both A and s for $A \in \mathbb{R}^+$.

Additionally, when considering the W_p metric as opposed to just the squared W_2 distance, we note that there are additional properties that are attractive. Many metrics can be defined on the space of probability measures, but the Wasserstein Distance has these key properties, as described in [4]:

- As mentioned earlier, the W_p distances are proper distances, meaning that they satisfy the definition of being a metric.
- Wasserstein Distances incorporate the geometry of the ground space χ, meaning if X and Y have weight 1 at points $x, y \in \chi$, then $W_p(X, Y)$ is simply the distance between x and y in χ. This property mirrors human intuition, meaning it matches the human perception of whether images are similar or not.
- Convergence of X_n to X in Wasserstein distance is equivalent to convergence in distribution, supplemented with $\mathbb{E}||X_n||^p \to \mathbb{E}||X||^p$, which is useful for proving central limit theorem type results. For instance, we have the following property for any real number a:

  $$W_p(aX, aY) = |a| W_p(X, Y) \tag{12}$$

  Additionally, we have the following property involving any fixed vector $x \in \chi$:

  $$W_p(X + x, Y + x) = W_p(X, Y) \tag{13}$$

- Since the Wasserstein distance is a solution to a minimization problem, it is easy to bound from above. Any joint distribution with the correct marginals provides an upper bound for the Wasserstein Distance.

2.5 The Wasserstein Barycenter

The Wasserstein Distance can also be used to define a notion of average in the form of a barycenter. Suppose we have a set of distributions $P_1, \ldots, P_N$ and we wanted to summarize these distributions with one "average" distribution. Taking the average of the distributions is one option, but the resulting average will not resemble any of the individual distributions (see Figure 1).

Instead, we use the Wasserstein Barycenter, which is the distribution P that minimizes

$$\sum_{j=1}^{N} W(P, P_j) \tag{14}$$

An example of this is shown in Figure 1. It's clear that because the Wasserstein metric incorporates the geometry of ground space (i.e. it mirrors human intuition of distance), it does a much better job of representing an "average" of these distributions.

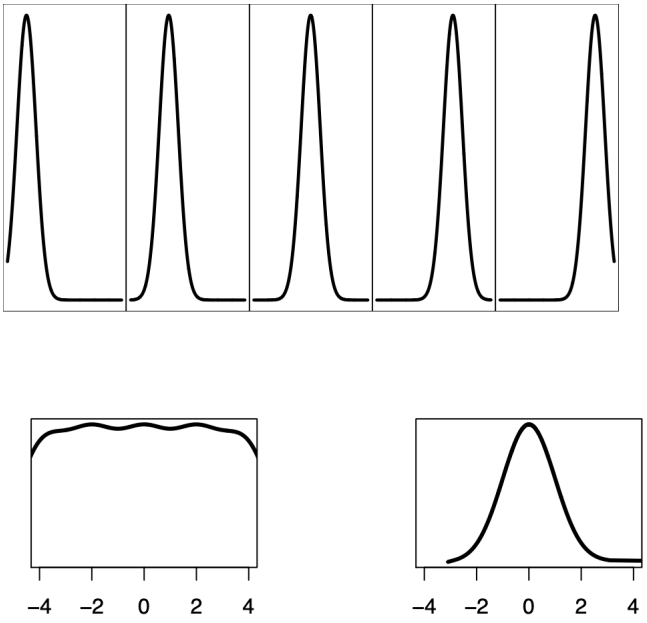

Figure 1: The top shows five distributions $P_1, \ldots, P_5$. Bottom left is the Euclidean average of these distributions, and bottom right shows the Wasserstein Barycenter

The same logic can be applied to datasets rather than distributions because we can regard a dataset as an empirical distribution.

3. The Kantorovich-Rubinstein Duality Theorem

3.1 The Duality Principle

We will now apply the Duality Principle to the Wasserstein Distance, which is the principle that any mathematical optimization problem can be viewed in one of two ways: primal or dual. What we have seen thus far is a primal optimization problem. We now instead turn to the dual problem, which provides a lower bound to the solution of the primal problem.

3.2 The Kantorovich-Rubinstein Norm

Given the general form of the Wasserstein distance in equation (7), we now look into the particular case where we have $p = 1$ and $d(x, y) = |x - y|$, also known as the W_1 distance metric. It can be shown that the W_1 distance

$$W_1(\mu, v) = \inf_{J \in \Gamma(\mu, v)} \int ||x - y|| dJ(x, y) \tag{15}$$

can be rewritten into a maximization problem with the form

$$W_1(\mu, v) = \sup \left\{ \int f(x) d\mu(x) - \int f(x) dv(x) : f \in F \right\} \tag{16}$$

where F denotes all maps from $\mathbb{R}^d$ to $\mathbb{R}$ such that $|f(y) - f(x)| = ||x - y||$ for all x and y. This form, called the Kantorovich-Rubinstein Duality norm, is the dual problem to the original primal problem. For a rigorous proof of this equivalence, please refer to [1].

Intuitively the Kantorovich-Rubinstein Duality can be understood by imagining that one is an industrialist trying to transfer coal from mines to factories. One way to go about this is to rent trucks to transport the coal, where the cost of transporting each ton of coal is $c(x, y)$ to go from location x to y. Alternatively, this can be thought of as having a cost $\varphi(x)$ to load all the coal up and a cost $\psi(y)$ to unload all the coal at the destination. Then, the cost would be the sum of $\varphi(x)$ and $\psi(y)$, and we have the following:

$$\varphi(x) + \psi(y) \leq c(x, y) \tag{17}$$

4. Applications

There are currently many applications for the Wasserstein distance, particularily in statistical machine learning. A small list of these applications include the following:

- Image processing
- Dimension Reduction
- Fluid simulation
- Signal processing
- Full-Waveform Inversion (FWI)

This is by no means an exhaustive list, but it gives an overview of the numerous applications of the Wasserstein Distance. We will focus on a few of the aforementioned applications below.

4.1 Image Processing

The ideas presented in Optimal Transport Theory are being vastly implemented in various image processing techniques. One particular use case is known as the color transfer problem, in which one modifies a source image I_X so that its colors match the colors of a target image I_Y. This is denoted mathematically as $I_X : \Omega \subset \mathbb{Z}^2 \to \Sigma \subset \mathbb{R}^3$, where Ω is the regular pixel grid and Σ is the quantized 3D RGB color space.

The goal is to find a new image I_Z that has similar geometry to I_X, but similar color distribution to I_Y. In other words, we want to compute a transformation T such that $\forall x \in \Omega$, $I_Z(x) = T(I_X(x))$ where the color histogram of I_Z is close to I_Y. The color histogram can be estimated using the empirical distributions μ_X, μ_Y, μ_Z for I_X, I_Y, and I_Z, respectively.

Consider Figure 2 below, which gives an example of two input images I_X and I_Y. The top row of Figure 2 shows the original images, while the bottom row shows the 3-D color distribution, labeled $\mu_X and \mu_Y$, but as a 2-D projection. The third column shows the distribution μ_Z, which is the result of applying the transform T to I_X when T is computed using an Optimal Transport framework. In the third column, the resulting image has the geometry of I_X and the color distribution of I_Y.

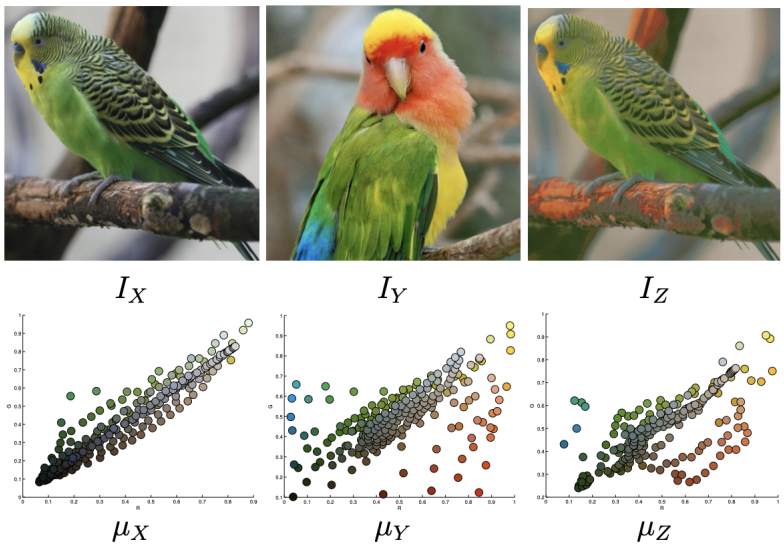

Figure 2: Example of color transfer using optimal transport.

In order to further improve color transformations, we can relax the mass conservation constraint that is imposed by Optimal Transport Theory. For details on how this is done, please refer to [5]. Below in Figure 3, we illustrate the effects of using a "relaxed" constraint, referred to as adaptive relaxation in this context. In Figure 3, I_X is the input image, I_Y is the desired color palette, and the result is shown in the rightmost image:

(a) I_X (b) I_Y (c) Adaptive model

Figure 3: Adaptive color transfer with relaxed Optimal Transport example

References

[1] Cedric Villani. *Topics in Optimal Transportation.* American Mathematical Society, Rhode Island, 2003

[2] Philippe Clement and Wolfgang Desch. *An Elementary Proof of the Triangle Inequality for the Wasserstein Metric.* Proceedings of the American Mathematical Society, 2007

[3] Srinath Mahankali and Dr. Yunan Yang. *Velocity Inversion Using the Quadratic Wasserstein Metric.* Courant Institute of Mathematical Sciences

[4] Victor M. Panaretos and Yoav Zemel. *Statistical Aspects of Wasserstein Distances.* Annual Review of Statistics and Its Applications, 2019

[5] Nicolas Papadakis. *Optimal Transport for Image Processing.* Signal and Image Processing. University de Bordeaux; Habilitation thesis, 2015