What is Wasserstein Distance?

The Wasserstein Distance, often called the Earth Mover's Distance (EMD), is a metric used to quantify the "cost" of transforming one probability distribution into another. Imagine you have two piles of sand (representing two distributions) of equal total volume. The Wasserstein Distance is the minimum amount of "work" required to move the first pile of sand to perfectly match the shape and location of the second pile. The "work" is calculated as the amount of sand moved multiplied by the distance it's moved.

This makes it particularly useful for:

Ordered categories or continuous data: When your data points have a natural order or numerical value (e.g., age groups, income brackets, time spent). Non-overlapping distributions: Unlike KL or JSD, Wasserstein distance provides a meaningful value even if the distributions don't share any common support. Sensitivity to shifts: It's sensitive to subtle shifts in probability mass, providing a more intuitive measure of how much one distribution needs to be "moved" to become the other.

Example: Comparing Student Performance Distributions

Let's say a school introduces a new teaching method. They want to compare the distribution of student grades (on a scale of 1-5, where 1 is lowest and 5 is highest) between students who followed the old method and those who followed the new method.

We'll define the categories as numerical values for clarity:

- Grade 1
- Grade 2
- Grade 3
- Grade 4
- Grade 5

Old Teaching Method (Distribution P):

- Grade 1: 10% (0.1)
- Grade 2: 20% (0.2)

- Grade 3: 40% (0.4)
- Grade 4: 20% (0.2)
- Grade 5: 10% (0.1)

New Teaching Method (Distribution Q):

- Grade 1: 5% (0.05)
- Grade 2: 15% (0.15)
- Grade 3: 30% (0.3)
- Grade 4: 35% (0.35)
- Grade 5: 15% (0.15)

To calculate the Wasserstein Distance, we need to think about how much "probability mass" (students) needs to be moved from one grade bin to another, and by how many "steps" (distance between grades).

Conceptual Calculation Steps:

Cumulative Distributions: It's often easiest to think about this in terms of cumulative distributions.

P (Cumulative):

- Grade 1: 0.1
- Grade 2: 0.1 + 0.2 = 0.3
- Grade 3: 0.3 + 0.4 = 0.7
- Grade 4: 0.7 + 0.2 = 0.9
- Grade 5: 0.9 + 0.1 = 1.0

Q (Cumulative):

- Grade 1: 0.05
- Grade 2: 0.05 + 0.15 = 0.2
- Grade 3: 0.2 + 0.3 = 0.5
- Grade 4: 0.5 + 0.35 = 0.85

- Grade 5: 0.85 + 0.15 = 1.0

Calculate the absolute difference between cumulative distributions at each point:

- Grade 1: $|0.1-0.05|=0.05$
- Grade 2: $|0.3-0.2|=0.1$
- Grade 3: $|0.7-0.5|=0.2$
- Grade 4: $|0.9-0.85|=0.05$

Sum these differences: The Wasserstein Distance is the sum of these absolute differences.

Wasserstein Distance = $0.05+0.1+0.2+0.05=0.4$

Interpretation:

- A Wasserstein Distance of 0.4 tells us that, on average, the "probability mass" (students) in the "Old Teaching Method" distribution would need to be shifted by 0.4 grade steps to match the "New Teaching Method" distribution.
    - A value of 0 would mean the distributions are identical.
    - A higher value means a greater "effort" or "distance" is required to transform one into the other.

In this example, the positive distance implies that the new teaching method has generally shifted student performance towards higher grades. The distance of 0.4 quantifies the average magnitude of this shift in a way that respects the ordering of the grade categories. This is a more intuitive result than what KL or JSD might give, especially if, for instance, the old method had no students in Grade 5 and the new method had some – KL/JSD would heavily penalize the zero probability, whereas Wasserstein would smoothly measure the shift from Grade 4 to Grade 5.

Common Real-Life Applications

1. **Machine Learning and Computer Vision**

   - **Generative Adversarial Networks (GANs):** Wasserstein GANs (WGANs) use the Wasserstein distance as the loss function instead of the traditional KL or Jensen-Shannon divergence. This helps to stabilize the training of GANs and prevent mode collapse, a common problem where the generator produces a limited variety of outputs.
   - **Image Processing:** In computer vision, it is used for tasks such as image retrieval and object recognition. The Wasserstein distance can be used to compare two images by treating them as probability distributions of pixel intensities. This allows for more robust image comparisons, even when images are slightly shifted or distorted.

2. **Statistics and Probability**

   - **Hypothesis Testing:** It can be used to perform non-parametric hypothesis tests to determine if two samples are drawn from the same distribution. The Wasserstein distance is particularly useful in this context because it is a true metric and can provide a more meaningful measure of distance between distributions.
   - **Data Drift Analysis:** It is used in data drift analysis to measure the difference between a reference distribution (e.g., training data) and a new, incoming distribution (e.g., production data). If the Wasserstein distance exceeds a certain threshold, it indicates that the data has drifted, and the model may need to be retrained.

3. **Other Applications**

   - **Optimal Transport:** The Wasserstein distance is closely related to the optimal transport problem, which seeks to find the most efficient way to move a pile of dirt from one location to another. This has applications in logistics, urban planning, and resource allocation. For example, it can be used to find the most efficient way to transport goods from a warehouse to multiple stores.