Probability

Hadar Tal

hadar.tal@mail.huji.ac.il

This paper is a summary of the educational materials and lectures from

- **Optimization for Computer Science** by Professor Tomer Koren, Tel Aviv University
- **Wikipedia**
- **3Blue1Brown** YouTube channel
- Introduction to the Wasserstein distance by "Applied Algebraic Topology Network" YouTube channel

Winter 2024

Contents

Chapter 1

Probability Measures

1.1 Probability Spaces

Definition 1.1.1. ***Probability Space***
A probability space is a triple $(\Omega, \mathcal{F}, P)$ *where:*

- Ω *is a set of outcomes.*
- $\mathcal{F}$ *is a sigma-algebra of events.*
- P *is a probability measure.*

Definition 1.1.2. ***Sigma-Algebra***
A sigma-algebra is a collection of subsets of a set Ω *that contains the empty set, is closed under complements, and is closed under countable unions.*

Definition 1.1.3. ***Probability Measure***
A probability measure is a function $P : \mathcal{F} \to [0, 1]$ *that satisfies:*

- $P(\Omega) = 1$.
- *For any countable collection of disjoint sets* $A_1, A_2, \ldots \in \mathcal{F}$,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Chapter 2

Distances and Metrics

2.1 Wasserstein Distance

The Wasserstein distance (also called the **kanterovich-rubinstein distance**, **optimal transport distance**, or **earth mover's distance**) is a measure of the distance between two probability distributions that has gained popularity in various fields, particularly in machine learning for its application in generative models.

While using function distance $\|f-g\|_\infty$ is a common way to measure the difference between two functions, it is not always the most appropriate.

Intuition:

- The Wasserstein distance can be thought of as the minimum "cost" required to transform one distribution into another, where the cost is measured by the amount of "mass" that must be moved times the distance it has to be moved.
- Amount of work needed to transform one distribution into another.
- Area between cumulative distribution functions (CDFs).
- Sensitive to outliers.

We will begin by looking at the discrete case of a Transport Plan.
Lets consider two discrete probability distributions $X = \{x_1, x_2, x_3, x_4\}$ and $Y = \{y_1, y_2, y_3\}$ with the probability mass functions:

$$P(X) = \{0.4, 0.1, 0.3, 0.2\} \quad \text{and} \quad P(Y) = \{0.3, 0.4, 0.3\}.$$

We can define a transport plan as a matrix π where π_{ij} represents the amount of mass that is transported from x_i to y_j.

	x_1	x_2	x_3	x_4
y_1	0.3	0.3	0	0
y_2	0.4	0.1	0.2	0
y_3	0.3	0	0.1	0.2

$$d_{W_1}\left(\sum_i \alpha_i \delta_{x_i}, \sum_j \beta_j \delta_{y_j}\right) = \min\{\sum_{i,j} \pi_{ij} d(x_i, y_j) : \pi_{ij} \geq 0, \sum_i \pi_{ij} = \beta_j, \sum_j \pi_{ij} = \alpha_i\}$$

Definition 2.1.1. ***Wasserstein Distance***
Given two probability measures μ and ν on a metric space (M, d), the Wasserstein distance of order p between μ and ν is defined as:

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{M \times M} d(x, y)^p \, d\gamma(x, y) \right)^{\frac{1}{p}},$$

where $\Gamma(\mu, \nu)$ denotes the collection of all measures on $M \times M$ with marginals μ and ν on the first and second factors, respectively.

And for W_1:

$$W_1(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{M \times M} d(x, y) \, d\gamma(x, y).$$

Example: Consider two one-dimensional distributions, μ which is a delta function at 0, and ν which is a delta function at 1. The Wasserstein distance of order 1 between μ and ν is simply the distance between the two points, which is 1.
This metric is particularly useful when comparing probability distributions that might not have the same support or when the cost of transportation is significant.

2.2 Total Variation Distance

2.3 Kullback-Leibler Divergence

2.4 Hellinger Distance