A Pot-Pourri of Probability & Optimal Transport

Davi Sales Barreira

September 9, 2020

Contents

Abstract

The main goal of these notes is to present an introduction to the transport inequalities, which consist of methods of using Optimal Transport Theory to obtain concentration inequalities in high-dimensional probability. Along the way, as new concepts are presented, some side-lining will be done, with the aim to explore some of these new concepts, before diving back in the proof of the inequalities.

Notation

- $P(\mathcal{X})$ - S Most of the content regarding transportation inequality is from

1 Introduction

Given two probability distributions μ, ν, there are many situations where one is interested in defining a way of measuring the distance between them. The Wasserstein distance is a metric that arises from the idea of optimal transport, and which has being gaining attention in Statistics and Machine Learning. One prominent example is the so called Wasserstein Generative Adversarial Network (WGAN), which uses this metric to evaluate how well the model generated distribution approximates the "real" distribution of the data. As will be shown shortly, the Wasserstein metric has several advantages compared to other metrics.

There are several ways of defining distances between two probability measures. Let's assume that ν, μ are defined on $(\Omega, \mathcal{F})$ and that $\nu \ll \lambda$, $\mu \ll \lambda$, for λ representing the Lebesgue measure. Below we present some example of distances:

$$\text{Total Variation}: \quad ||\mu - \nu||_{TV} = \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)| = \frac{1}{2} \int_\Omega \left| \frac{d\mu}{d\lambda} - \frac{d\nu}{d\lambda} \right| d\lambda$$

$$\text{Hellinger}: \quad \sqrt{\int_\Omega \left(\sqrt{\frac{d\mu}{d\lambda}} - \sqrt{\frac{d\nu}{d\lambda}} \right)^2 d\lambda}$$

$$L_2: \quad \int_\Omega \left(\frac{d\mu}{d\lambda} - \frac{d\nu}{d\lambda} \right)^2 d\lambda$$

As pointed out by Wasserman [2018] in his lecture notes, although such distances are useful, there are drawbacks:

- If one distribution is discrete and the other is continuous, they cannot be compared. If $X \sim U(0,1)$ and Y is uniform on $\{0, 1/N, 2/N, ..., 1\}$, although this distributions are very similar, their total variation is 1 (which is the maximum value). The Wasserstein distance is $1/N$, which is reasonable.
- These distances ignore the "geometry of the underlying space", while the Wasserstein distance preserves it, as shown in Figure 1.
- When "averaging" different distributions, one might be interested in obtaining a similar distribution, avoiding smoothing. This can be done using the Wasserstein barycenter, as shown in Figure 2.

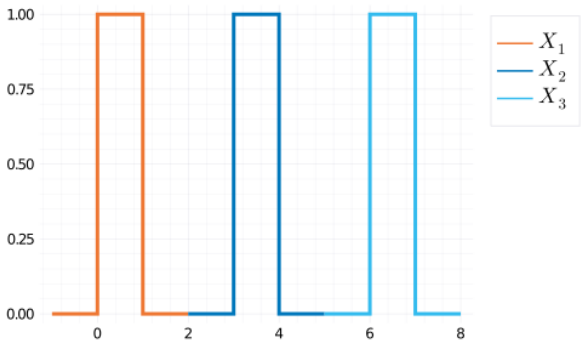

Figure 1: Each pair has the same distance in L_2, Hellinger and TV. But using Wasserstein, X_1 is closer to X_2 than to X_3.

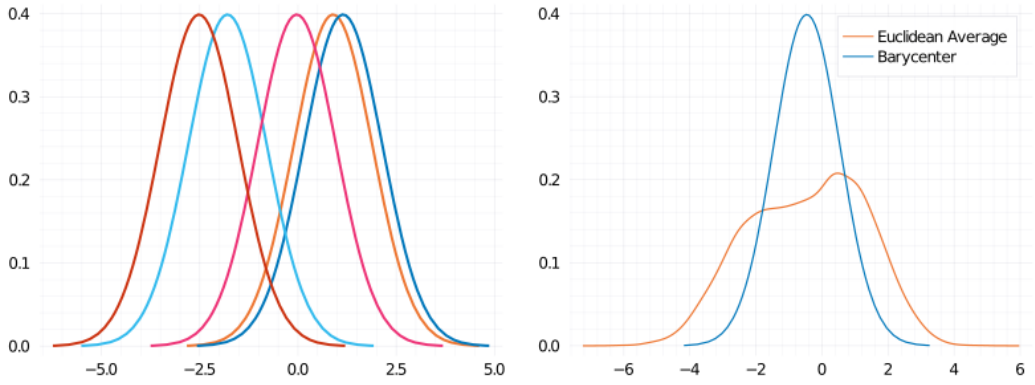

Figure 2: In the left you have different distributions, and in the right, there is a comparison between averaging these distribution versus finding the Wasserstein barycenter.

References

Larry Wasserman. Statistical methods for machine learning - lecture notes, 2018.