Notes on Optimal Transport
ENSAE
3A

Gabriel Romon

Last updated: Wednesday 6th February, 2019 at 15:32

1 Introduction

These are notes I gathered from a seminar session given by Shuangjian Zhang at ENSAE and from lectures given by Marco Cuturi at MLSS 2019 in South Africa.

Likelihood maximization is an instance of generative modeling: given data points $x_1, \dots, x_N \in \mathbb{R}^d$ and a family of densities $(f_\theta)_{\theta \in \Theta}$, we look for the f_θ that matches the most the empirical data distribution defined as $\nu_{\text{data}} := \frac{1}{N} \sum_{i=1}^N \delta_{x_i}$. This is done by maximizing the log-likelihood $\frac{1}{N} \sum_{i=1}^N \log f_\theta(x_i)$. This quantity exists only if $f_\theta(x_i) > 0$ for all i, which forces the densities to have the whole of $\mathbb{R}^d$ as support. Likelihood maximization can be given a geometric interpretation: if one overlooks that ν_{data} and f_θ are not absolutely continuous with respect to the same measure (the former is discrete), minimizing the Kullback-Leibler divergence between the two writes as

$$\begin{aligned}\operatorname{argmin}_{\theta \in \Theta} \mathrm{KL}(\nu_{\text{data}} || f_\theta) &= \operatorname{argmin}_{\theta \in \Theta} E_{X \sim \nu_{\text{data}}} [\log(f_{\text{data}}(X)) - \log(f_\theta(X))] \\ &= \operatorname{argmin}_{\theta \in \Theta} -E_{X \sim \nu_{\text{data}}} \log(f_\theta(X)) \\ &= \operatorname{argmax}_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^N \log f_\theta(x_i)\end{aligned}$$

A weakness of likelihood maximization is its poor scalability to high-dimensional settings, which are commonplace. For instance, 100×100 images with 3 channels live in $\mathbb{R}^{30.000}$. Instead of working directly in the data space $\mathbb{R}^d$, we may rather consider a latent space $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p), \mu)$ with $p \ll d$, and measurable functions $g_\theta : \mathbb{R}^p \to \mathbb{R}^d$ (e.g deconvolution networks). We define the pushforward measure $g_{\theta\sharp}\mu$ by $\forall B \in \mathcal{B}(\mathbb{R}^d), g_{\theta\sharp}\mu(B) := \mu(g_\theta \in B)$ and we look for θ such that $g_{\theta\sharp}\mu$ matches ν_{data}. This requires setting a metric on the space of probability measures, and fortunately, many exist: Hellinger, Kantorovitch, MMD, Wasserstein. Some of these metrics arise from the theory of optimal transport.

2 Optimal transport