Problem Set 5: Due Tuesday Week 15

(1) Let f and g be two non-negative functions such that

$$\int_{\mathbb{R}^d} f(x)\,dx = \int_{\mathbb{R}^d} g(x)\,dx = 1.$$

Show that we always have the following inequality,

$$\int_{\mathbb{R}^d} f(x)\log(g(x))\,dx \leq \int_{\mathbb{R}^d} f(x)\log(f(x))\,dx,$$

with equality holding if and only if $f = g$ a.e. *Hint: Apply Jensen's inequality, for the right convex expression, to a properly chosen integral expression involving the quotient g/f.*

(2) Let $\mathcal{P}$ denote the set of probability measures μ in $\mathbb{R}^d$ such that

$$\int_{\mathbb{R}^d} |x|^2\,d\mu(x) < \infty$$

Given $\mu, \nu \in \mathbb{P}$, let $\Gamma(\mu,\nu)$ be the set of probability measures in $\mathbb{R}^d \times \mathbb{R}^d$ with firsts marginal equal to μ and second marginal equal to ν. Show that the Kantorovich problems with costs

$$c_1(x,y) = |x-y|^2$$
$$c_2(x,y) = -(x,y)$$

are related to one another.

(3) Given n different points $y_1,\ldots,y_n \in [0,1]$, let μ_y be the measure given by

$$\mu_y = \frac{1}{n}\sum_{k=1}^{n} \delta_{y_k}$$

Let μ_0 denote the uniform distribution over $[0,1]$. Determine a formula for the solution to the Kantorovich problem for the quadratic cost with source measure μ_0 and target measure μ_y.

(4) Consider the previous problem and take $n = 3$ and $n = 4$, and in each instance the set of n different points $\{y_1,\ldots,y_n\}$ in $[0,1]$, and find a configuration for which the (quadratic) cost of transporting μ_0 to μ_y is smallest possible (with that fixed n). Compare the optimal value between $n = 3$ and $n = 4$. What do you think happens in general?

(5) Consider $N = 150$ points in $\mathbb{R}^2$ sampled i.i.d. from a Gaussian mixture made out of three, equally weighted Gaussians. The Gaussians have covariance matrix given by the identiyy and with three different means: $(3,0), (0,-3)$ and $(0,3)$. Starting from any initial partition in the 3 components, run the K-means algorithm over the set of 150 points with $K = 3$.