# Homework 3
# Optimal Transport, High-dimensional Regression

CMU 10-716: Advanced Machine Learning (Spring 2020)

OUT: March 24, 2020
DUE: **April 7, 2020, 3:00 PM**.

## Instructions:

- **Collaboration policy:** Collaboration on solving the homework is allowed, after you have thought about the problems on your own. It is also OK to get clarification (but not solutions) from books or online resources, again after you have thought about the problems on your own. There are two requirements: first, cite your collaborators fully and completely (e.g., "Bob explained to me what is asked in Question 4.3"). Second, write your solution *independently*: close the book and all of your notes, and send collaborators out of the room, so that the solution comes from you only.
- **Submitting your work:** Assignments should be submitted as PDFs using Gradescope unless explicitly stated otherwise. Each derivation/proof should be completed on a separate page. Submissions can be handwritten, but should be labeled and clearly legible. Else, submission can be written in LaTeX. Upon submission, label each question using the tempate provided by Gradescope.

# 1 Optimal Transport

ecall the Kantorovich optimal transport problem:

$$\mathcal{L}_c(\alpha, \beta) = \min_{\pi \in U(\alpha,\beta)} \int_{\mathbb{R}^2} c(x, y) d\pi(x, y),$$

where $c : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ is a given cost function, $\alpha, \beta \in \mathcal{M}(\mathbb{R})$ are two discrete probability measures, and $U(\alpha, \beta)$ is the set of all joint distributions $\pi$ with marginals $\alpha$ and $\beta$. Suppose that $\alpha, \beta$ have cumulative distributions $F$ and $G$ respectively. Assume $c(x, y) = d(x - y)$ where $d$ is strictly convex and continuous. Let $\pi^\dagger$ be the measure on $\mathbb{R}^2$ with cumulative distribution function $H(x, y) = \min\{F(x), G(y)\}$. In this problem, we will show that $\pi^\dagger$ is a solution to the Kantorovich optimal transport problem.

a) Let $\pi^* \in U(\alpha, \beta)$ be the solution to the Kantorovich optimal transport problem (you can assume such a solution exists with no proof). Show that if $(x_1, y_1), (x_2, y_2) \in \text{supp}(\pi^*)$ and $x_1 < x_2$ then $y_1 \leq y_2$.

Hint: You can use, without proving, that the $\text{supp}(\pi^*)$ is monotone, i.e. for $(x_1, y_1), (x_2, y_2) \in \text{supp}(\pi^*)$

$$d(x_1 - y_1) + d(x_2 - y_2) \leq d(x_1 - y_2) + d(x_2 - y_1)$$

b) Show that $\pi^\dagger = \pi^*$. In particular, show that $\pi^* ((-\infty, x], (-\infty, y]) = \min\{F(x), G(y)\}$.

Hint: Consider the sets $A = (-\infty, x] \times (y, +\infty)$ and $B = (x, +\infty) \times (-\infty, y]$ and what can be said about $\pi^*(A)$ and $\pi^*(B)$.

# 2 Square Root Lasso

The square-root Lasso is given by

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{\sqrt{n}} ||y - \mathbf{X}\theta||_2 + \gamma_n ||\theta||_1 \right\}.$$

1. (5 pts) Show that any square-root Lasso estimate $\hat{\theta}$ satisfies the equality

$$\frac{\frac{1}{n}\mathbf{X}^T \left( \mathbf{X}\hat{\theta} - y \right)}{\frac{1}{\sqrt{n}} ||y - \mathbf{X}\hat{\theta}||_2} + \gamma_n \hat{z} = 0$$

where $\hat{z} \in \mathbb{R}^d$ belongs to the subdifferential of the $\ell_1$-norm at $\hat{\theta}$.

2. (5 pts) Suppose $y = \mathbf{X}\theta^* + w$ where the unknown regression vector $\theta^*$ is $s$-sparse. Let $S$ be the set of indices where $\theta^*$ is non-zero. Use part (1) to establish that the error $\widehat{\Delta} = \hat{\theta} - \theta^*$ satisfies the basic inequality

$$\frac{1}{n} ||\mathbf{X}\widehat{\Delta}||_2^2 \leq \langle \widehat{\Delta}, \frac{1}{n} \mathbf{X}^T w \rangle + \gamma_n \frac{||y - \mathbf{X}\hat{\theta}||_2}{\sqrt{n}} \{ ||\widehat{\Delta}_S||_1 - ||\widehat{\Delta}_{S^c}||_1 \}.$$

**Hints:**

(a) you can use without proof that since $\hat{z}$ is a sub-differential of $L_1$ norm at $\hat{\theta}$, $\hat{\Delta}^T\hat{z} \geq ||\hat{\theta}||_1 - ||\theta^*||_1$.

(b) you will need to show that $||\theta^*||_1 - ||\hat{\theta}||_1 \leq ||\hat{\Delta}_S||_1 - ||\hat{\Delta}_{S^c}||_1$.

3. (10 pts) Suppose that in addition $\mathbf{X}$ satisfies the RE condition with parameter $\kappa$ so that

$$\frac{1}{n}||\mathbf{X}\hat{\Delta}||^2 \geq \kappa||\hat{\Delta}||^2.$$

Also, let $\gamma_n \geq 2\frac{\|X^T w\|_\infty}{\sqrt{n}\|w\|_2}$ and suppose that $\kappa - \gamma_n^2 s \geq \rho$ for some constant $\rho > 0$. Show that there is a constant $c$ such that

$$||\widehat{\theta} - \theta^*||_2 \leq c\frac{||w||_2}{\sqrt{n}}\gamma_n\sqrt{s}.$$

**Hints:**

(a) Use Holder's inequality.

(b) Use the fact that for a $d$-dimensional vector $||\theta||_1 \leq \sqrt{d}||\theta||_2$.

# 3 L-Infinity Lasso

Consider the sparse linear model $y = \mathbf{X}\theta^* + w$, where $\mathbf{X} \in \mathbb{R}^{n\times d}$ is fixed, $w_i$ $(i = 1, \ldots, n)$ are sampled iid. from $\mathcal{N}(0, \sigma^2)$ independently of everything else, and $\theta^* \in \mathbb{R}^d$ is supported on a subset $S$. Suppose that the sample covariance matrix $\hat{\Sigma} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ has its diagonal entries uniformly upper bounded by one, and that for some parameter $\gamma > 0$, it also satisfies an $\ell_\infty$-curvature condition of the form

$$\|\hat{\Sigma}\Delta\|_\infty \geq \gamma\|\Delta\|_\infty \quad \forall\ \Delta \in \mathcal{C}_3(S),$$

where $\mathcal{C}_3(S) = \{\Delta \in \mathbb{R}^d \mid \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1\}$. Recall that the Lasso solution $\hat{\theta}$ is defined as

$$\hat{\theta} = \arg\min_{\theta\in\mathbb{R}^d}\left\{\frac{1}{2n}\|y - \mathbf{X}\theta\|_2^2 + \lambda_n\|\theta\|_1\right\}.$$

In this problem, we will in steps, show that with the regularization parameter $\lambda_n = 4\sigma\sqrt{\frac{\log d}{n}}$, any Lasso solution satisfies the following $\ell_\infty$-bound,

$$\|\hat{\theta} - \theta^*\|_\infty \leq \frac{6\sigma}{\gamma}\sqrt{\frac{\log d}{n}},$$

with high probability.

1. First prove that, if $\lambda_n \geq 2\left\|\frac{X^T w}{n}\right\|_\infty$, then:

$$\hat{\Delta} = \hat{\theta} - \theta \in \mathcal{C}_3(S)$$

Hint: The regularized empirical risk of $\hat{\theta}$ is smaller than that of $\theta^*$.

2. Next, by using that 0 belongs to the subgradient of the loss at $\hat{\theta}$, we may obtain that for some $z$ with $\|z\|_\infty \leq 1$:

$$\frac{X^T X\hat{\theta} - X^T y}{n} + \lambda_n z = 0$$

Prove then that:

$$\|\hat{\Delta}\|_\infty \leq \frac{3}{2\gamma}\lambda_n$$

3. Finally, we know from maximal inequality that:

$$\Pr(\frac{\|w^T \mathbf{X}\|_\infty}{n} \geq t) \leq 2de^{-\frac{nt^2}{2\sigma^2}} \leq \frac{2}{d}$$

for $t = 2\sigma\sqrt{\frac{\log d}{n}}$.

Conclude using parts 1 and 2 that with probability at least $1 - 2/d$:

$$\|\hat{\theta} - \theta^*\|_\infty \leq \frac{6\sigma}{\gamma}\sqrt{\frac{\log d}{n}}$$