Exercise 2: Robustness and Lasso (5 pts)

Consider the linear regression problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \|X\mathbf{w}-\mathbf{y}\|_2, \tag{3}$$

where $X \in \mathbb{R}^{n\times d}$, $\mathbf{y} \in \mathbb{R}^n$, n is the number of training examples and d is the number of features. For simplicity we omitted the bias. Now suppose we perturb each *feature* independently, and we are interested in solving the robust linear regression problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \max_{\forall j, \|\mathbf{z}_j\|_2\le\lambda} \|(X+Z)\mathbf{w}-\mathbf{y}\|_2, \tag{4}$$

where the perturbation matrix $Z = [\mathbf{z}_1, \ldots, \mathbf{z}_d] \in \mathbb{R}^{n\times d}$. Prove that robust linear regression is exactly equivalent to (square-root) Lasso (note the absence of the square on the ℓ_2 norm):

'='

$$\min_{\mathbf{w}\in\mathbb{R}^d} \|X\mathbf{w}-\mathbf{y}\|_2 + \lambda\|\mathbf{w}\|_1, \tag{5}$$

where recall that $\|\mathbf{w}\|_1 = \sum_j |w_j|$.

[Hint: Start from (4), apply the Cauchy-Schwarz inequality $\|\mathbf{w}\|_2 = \max_{\|\mathbf{u}\|_2\le 1} \mathbf{w}^\top\mathbf{u}$ a few times and a few other tricks, in order to convert it into (5).]

n : # training examples.

d : # features

Perturbation:

$Z = [\vec{z}_1, \ldots \vec{z}_d] \in \mathbb{R}^{n\times d}$

$X \in \mathbb{R}^{n\times d}$ $\quad \vec{y} \in \mathbb{R}^n$

$$\boxed{\min_{\vec{w}\in\mathbb{R}^d} \max_{\forall j, \|\vec{z}_j\|_2<\lambda} \|(X+Z)\vec{w}-\vec{y}\|_2} \overset{?}{=} \boxed{\min_{\vec{w}\in\mathbb{R}^d} \|X\vec{w}-\vec{y}\|_2 + \lambda\|\vec{w}\|_1}$$

$$\max_{\forall j, \|\vec{z}_j\|_2<\lambda} \|(X+Z)\vec{w}-\vec{y}\|_2 = \max_{\forall j, \|\vec{z}_j\|<\lambda} \|X\vec{w}-\vec{y}+Z\vec{w}\|_2$$

(Triangle Inequality) $\Rightarrow$

$$\le \underbrace{\max_{\forall j, \|\vec{z}_j\|<\lambda} \|X\vec{w}-\vec{y}\|_2}_{=\|X\vec{w}-\vec{y}\|_2} + \max_{\forall j, \|\vec{z}_j\|<\lambda} \|Z\vec{w}\|_2$$

$[\ \ ][\ ] = [\circ]$

$$= \max_{\forall j, \|\vec{z}_j\|\le\lambda} \sum_{i=1}^{d} \|\vec{z}_i w_i\|_2 \le \sum_{i=1}^{d} \lambda\|w_i\|_2 = \sum_{i=1}^{d} \lambda|w_i| = \lambda\|\vec{w}\|_1$$

$$\le \|X\vec{w}-\vec{y}\|_2 + \lambda\|\vec{w}\|_1$$

As a result:

$$\therefore \max_{\forall j, \|\vec{z}_j\|_2<\lambda} \|(X+Z)\vec{w}-\vec{y}\|_2 \le \|X\vec{w}-\vec{y}\|_2 + \lambda\|\vec{w}\|_1$$

$$\boxed{\therefore \min_{\vec{w}\in\mathbb{R}^d} \max_{\forall j, \|\vec{z}_j\|_2<\lambda} \|(X+Z)\vec{w}-\vec{y}\|_2 \equiv \min_{\vec{w}\in\mathbb{R}^d} \|X\vec{w}-\vec{y}\|_2 + \lambda\|\vec{w}\|_1}$$

QED