# Homework 3
# Uniform Laws, Sparse Linear Models

CMU 10-716: Advanced Machine Learning (Spring 2019)

OUT: Feb. 27, 2019
DUE: **March 8, 2019, 11:59 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.
- **Start Early.**

# 1 Rademacher Complexity [27 pts]

1. [**2 pts each**] Recall the definition of Rademacher complexity of a bounded set $\mathcal{S} \subset \mathbb{R}^d$:

$$\mathcal{R}_d(\mathcal{S}) := \mathbb{E}\left[\sup_{s\in\mathcal{S}} \left|\frac{1}{d}\sum_{i=1}^{d} \varepsilon_i s_i\right|\right].$$

Let's further introduce the notation:

$$\mathcal{R}_d^\circ(\mathcal{S}) := \mathbb{E}\left[\sup_{s\in\mathcal{S}} \frac{1}{d}\sum_{i=1}^{d} \varepsilon_i s_i\right],$$

where $s_i$ is the $i$th element of $s$ and $\varepsilon_i$ are independent Rademacher RVs. Answer whether the following are True or False with a few sentences of explanations for each.

(a) $\mathcal{R}_d^\circ(\mathcal{S}) \le \mathcal{R}_d(\mathcal{S}) = \mathcal{R}_d^\circ(\mathcal{S} \cup -\mathcal{S})$
(b) $\forall u \in \mathbb{R}^d, \mathcal{R}_d^\circ(\mathcal{S}+u) = \mathcal{R}_d^\circ(\mathcal{S})$
(c) $\mathcal{R}_d^\circ(\mathcal{S}_1 \cup \mathcal{S}_2) \le \mathcal{R}_d^\circ(\mathcal{S}_1) + \mathcal{R}_d^\circ(\mathcal{S}_2)$
(d) $\mathcal{R}_d^\circ(\mathcal{S}_1 + \mathcal{S}_2) = \mathcal{R}_d^\circ(\mathcal{S}_1) + \mathcal{R}_d^\circ(\mathcal{S}_2)$
(e) $\mathcal{R}_d^\circ(c\mathcal{S}) = |c|\mathcal{R}_d^\circ(\mathcal{S})$
(f) $\mathcal{R}_d^\circ(\mathcal{S}) = \mathbb{E}\Big[\sup_{s,s'\in\mathcal{S}} \frac{1}{2d}\sum_{i=1}^{d} \varepsilon_i(s_i - s_i')\Big]$

2. **[15 pts]** Suppose $\mathcal{F}$ is a set of functions $f : \mathcal{Z} \to [0,1]$. Suppose $Z_1, Z_2, \cdots, Z_n$ are i.i.d. samples from a distribution $P$. Consider the following:

$$\Delta_n(Z_1, Z_2, \cdots, Z_n) = \|P_n - P\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n}(f(Z_i) - \mathbb{E}[f(Z)])\right|$$

and

$$\mathcal{R}_n(\mathcal{F}(Z_1, Z_2, \cdots, Z_n)) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(Z_i)\right|\right],$$

where $\varepsilon_i$s are independent Rademacher RVs independent of $Z_1, Z_2, \cdots, Z_n$.

Show the following:

(a)

$$\mathbb{P}\Big(\Delta_n(Z_{1:n}) - 2\mathcal{R}_n(\mathcal{F}(Z_{1:n})) \ge \mathbb{E}\left[\Delta_n(Z_{1:n}) - 2\mathcal{R}_n(\mathcal{F}(Z_{1:n}))\right] + t\Big) \le e^{-2nt^2/25} \quad \forall\, t > 0.$$

(b)

$$\mathbb{P}\Big(\Delta_n(Z_{1:n}) - 2\mathcal{R}_n(\mathcal{F}(Z_{1:n})) \ge t\Big) \le e^{-2nt^2/25} \quad \forall\, t > 0.$$

# 2 Regularized Least Squares [25 pts]

Consider a linear model $Y = X^T\beta + \epsilon,\ X, \beta \in \mathbb{R}^d,\ Y, \epsilon \in \mathbb{R}$. Given samples $\{(X_i, Y_i)\}_{i=1}^n$, $\beta$ can be estimated by minimizing a regularized *mean squared error* loss function as follows.

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n}\sum_{i}(Y_i - X_i^T\beta)^2 + \lambda R(\beta).$$

Assume $\boldsymbol{X}^T\boldsymbol{X} = I$, where the $i$th row of $\boldsymbol{X}$ is $X_i$. Find the explicit form of $\hat{\beta}$ for the following three cases:

1. **[10 pts]** $R(\beta) = ||\beta||_0$
2. **[10 pts]** $R(\beta) = ||\beta||_1$
3. **[5 pts]** $R(\beta) = ||\beta||_2^2$

Hint: To solve part 1 and 2, expand the loss functions using the components of $\beta$, and minimize w.r.t. each component.

# 3 $\ell_\infty$ - bounds for the Lasso [18 pts]

Consider the sparse linear model $y = \boldsymbol{X}\theta^* + w$, where $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ 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 $\hat{\Sigma} = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}$ 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

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

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

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

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.

**Hints:**

1. First use the fact that $\hat{\theta}$ yields a smaller loss than $\theta^*$. Rearrange the resulting expression to yield a basic inequality. Then show that $\lambda_n \geq 2\left\|\frac{X^T w}{n}\right\|_\infty$ implies that $\hat{\Delta} = \hat{\theta} - \theta \in \mathbb{C}_3(S)$.
2. Use the fact that $\hat{\theta}$ yields the optimal solution and therefore $\mathbf{0}$ belongs to the subgradient of the loss at $\hat{\theta}$. Then rearrange the expression and apply the $\|\cdot\|_\infty$ norm on both sides of the equality.
3. Finally, use the $\ell_\infty$-curvature condition and that $\lambda_n \geq 2\left\|\frac{X^T w}{n}\right\|_\infty$ to show that $\|\hat{\Delta}\|_\infty \leq \frac{3}{2\gamma}\lambda_n$. Conclude the proof by showing $\lambda_n \geq 2\left\|\frac{X^T w}{n}\right\|_\infty$ with high probability.