Homework 4
High-dimensional models, lower bounds, statistical complexity of optimization

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

OUT: Mar. 22, 2019
DUE: **April 4, 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 template provided by Gradescope.
- **Start Early.**

1 Square-root Lasso [20 pts]

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. (2 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 ℓ_1-norm at $\hat{\theta}$.

2. (2 pts) Suppose $y = \mathbf{X}\theta^* + w$ where the unknown regression vector θ^* is S-sparse. Use part (1) to establish that the error $\widehat{\Delta}$ satisfies the basic inequality

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

3. (8 pts) Suppose that $\gamma_n \geq 2\frac{||\mathbf{X}^T w||_\infty}{\sqrt{n}||w||_2}$. Show that the error vector satisfies the cone constraint $||\widehat{\Delta}_{S^c}||_1 \leq 3||\widehat{\Delta}_S||_1$. The significance of square-root lasso is that γ_n is not dependent on the scale of the noise w due to the normalization term $\|w\|_2$.

   Hints:

   (a) Norms $||\cdot||$ are convex.

   (b) First-order characterization of convex functions: if f is differentiable, then f is convex if and only if $\text{dom}(f)$ is convex and $f(y) \geq f(x) + \nabla f(x)^T(y-x)$ for all $x, y \in \text{dom}(f)$.

   (c) Start the proof with bounding $||y - \mathbf{X}\widehat{\theta}||_2 - ||y - \mathbf{X}\theta^*||_2$ from above and below.

   (d) Notice that by the optimality of $\widehat{\theta}$, we have $||y-\mathbf{X}\widehat{\theta}||_2 \leq ||w||_2 + \sqrt{n}\gamma_n \left(||\widehat{\Delta}_S||_1 - ||\widehat{\Delta}_{S^c}||_1\right)$.

4. (8 pts) Suppose in addition that $\mathbf{X}$ satisfies an RE condition over S with parameter (κ, α) and $\kappa - \gamma_n^2 s \geq \rho$ 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) Using part (d) of the above hints, we can get $||y - \mathbf{X}\widehat{\theta}||_2 \leq ||w||_2 + \sqrt{n}\gamma_n||\widehat{\Delta}_S||_1$.

2 Lower Bound [25 pts each]

1. Let $X_1, \cdots, X_n \sim P$ where $X_i \in \mathbb{R}$. Assume $P = \text{Uniform}(0, \theta)$ where $0 < \theta < M$. Given a set of distributions $\mathcal{P} = \{\text{Uniform}(0,\theta) : \theta \in \mathbb{R}\,\text{and}\, 0 < \theta < M\}$, out goal is to estimate θ. Define the minimax risk (under ℓ_1 loss) for estimating θ as follows

$$R_n = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P|\hat{\theta} - \theta|.$$

   Show that $R_n \asymp \frac{1}{n}$, which means

$$\frac{c}{n} \leq R_n \leq \frac{C}{n}$$

   for some constants $0 < c \leq C \leq \infty$. Please enumerate at least one pair of $\{c, C\}$.

   Hints:

   (a) For the upper bound, you can use the estimator $\hat{\theta}$ as the maximum of $\{X_1, \cdots, X_n\}$. You can use the following fact as given that:

$$\frac{\hat{\theta}}{\theta} \sim \text{Beta}(n, 1).$$

   (b) For the lower bound, consider Le Cam's method.

2. Let $X_1, \cdots, X_n \sim P$ where $X_i = (X_i^1, \cdots, X_i^d) \in \mathbb{R}^d$ with $d \geq 2$. Assume $P = N(\theta, I)$ where $\theta = (\theta^1, \cdots, \theta^d)$ and I is the $d \times d$ identity matrix. Given a set of distributions $\mathcal{P} = \{N(\theta, I) : \theta \in \mathbb{R}^d\}$, our goal is to estimate θ. Define the minimax risk (under ℓ_∞ loss) for estimating θ as follows

$$R_n = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \|\hat{\theta} - \theta\|_\infty.$$

Show that $R_n \asymp \sqrt{\frac{\log d}{n}}$, which means

$$c\sqrt{\frac{\log d}{n}} \leq R_n \leq C\sqrt{\frac{\log d}{n}}$$

for some constants $0 < c \leq C \leq \infty$. Please enumerate at least one pair of $\{c, C\}$.

Hints:

(a) For the upper bound, you can use the estimator $\hat{\theta}$ as the mean of $\{X_1, \cdots, X_n\}$. Then, show that: $\hat{\theta} - \theta \sim N(0, \frac{1}{n}I)$. Then use the maximal inequality for subgaussian R.V.:
If $X_1, \cdots, X_n$ are n R.V.s s.t. X_i are sub-Gaussian with parameter σ, then

$$\mathbb{E}[\max_{1 \leq i \leq n} |X_i|] \leq \sigma\sqrt{2\text{log}(2n)}.$$

(b) For the lower bound, consider the "generalized Fano's method" discussed in class, also summarized below:
Let $F = \{P_1, \cdots, P_M\} \subset \mathcal{P}$. Let $\theta(P)$ be a parameter taking values in a metric space with metric d. Then

$$R_n \geq \frac{s}{2}\Big(1 - \frac{n\beta + \text{log}2}{\text{log}M}\Big)$$

where

$$s = \min_{j \neq k} d\big(\theta(P_j), \theta(P_k)\big)$$

and

$$\beta = \max_{j \neq k} KL(P_j, P_k).$$

3 Statistical complexity of optimization [15 pts]

Let us consider a space of input-output pairs $(x, y) \in X \times Y$. The discrepancy between the predicted output $\hat{y}$ and the real output y is measured with a loss function $l(\hat{y}, y)$. Our benchmark is the function f^* that minimizes the expected risk. Let us define

$$E(f) = \mathbb{E}[l(f(x), y)],$$

and the empirical risk as

$$E_n(f) = \frac{1}{n}\sum_{i=1}^{n} l(f(x_i), y_i) = \mathbb{E}_n[l(f(x), y)].$$

Let us choose a family $\mathcal{F}$ of candidate prediction functions and we try to find the function that minimizes the empirical risk

$$f_n = \text{arg min}_{f \in \mathcal{F}} \mathbb{E}_n(f).$$

Since the optimal function f^* is unlikely to belong to the family $\mathcal{F}$, we also define

$$f^*_{\mathcal{F}} = \text{arg min}_{f \in \mathcal{F}} \mathbb{E}(f).$$

Let us assume that f^*, $f^*_{\mathcal{F}}$ and f_n are well defined and unique. If our minimization algorithm returns an approximate solution $\tilde{f}_n$ such that $E_n(\tilde{f}_n) < E_n(f_n) + \rho$, We can write

$$\begin{aligned}
\mathcal{E} &= \mathbb{E}[E(\tilde{f}_n) - E(f^*)] \\
&= \mathbb{E}[E(f^*_{\mathcal{F}}) - E(f^*)] + \mathbb{E}[E(f_n) - E(f^*_{\mathcal{F}})] + \mathbb{E}[E(\tilde{f}_n) - E(f_n)] \\
&= \mathcal{E}_{app} + \mathcal{E}_{est} + \mathcal{E}_{opt}.
\end{aligned}$$

Let $\mathcal{F}$ be the class of binary linear classifiers $\text{sign}(w^T x)$ pamameterized by $w \in \mathbb{R}^d$, and $l(\cdot, y)$ be 1-Lipshcitz for all $y \in Y$. Also assume that the absolute value of $l(\cdot, \cdot)$ is upper bounded by 1. Show that for some positive constant c,

$$\mathcal{E}_{est} + \mathcal{E}_{opt} \leq c\sqrt{\frac{d \log(n+1)}{n}} + \rho$$

Hint: You may use the following facts without proof:

- For $|g(X)| \leq 1$, $|\mathcal{R}_n(\mathcal{F} + g) - \mathcal{R}_n(\mathcal{F})| \leq \sqrt{2 \log 2/n}$
- Consider $\phi : \mathbb{R} \times \mathcal{Z} \rightarrow \mathbb{R}$.
  Suppose, $\forall z, \alpha \rightarrow \phi(\alpha, z)$ is 1-Lipschitz and $\phi(0, z) = 0$.
  Define $\phi(\mathcal{F}) = \{z \rightarrow \phi(f(z), z) : f \in \mathcal{F}\}$. Then $\mathcal{R}_n(\phi(\mathcal{F})) \leq 2\mathcal{R}_n(\mathcal{F})$.