Homework 5
Computation Complexity

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

OUT: April 8, 2019
DUE: **April 24, 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 Stability and Generalization [20 pts each]

1. **(20pts)** Consider a set of samples $S = \{(x_i, y_i)\}_{i=1}^m \in (\mathcal{X} \times \mathcal{Y})^m = D^m$ and a function class $\mathcal{F}$ of functions $f : \mathcal{X} \to \hat{\mathcal{Y}}$ where $\mathcal{Y}, \hat{\mathcal{Y}} \in \mathbb{R}$. We also have the loss function $\ell : \hat{\mathcal{Y}} \times \mathcal{Y} \to [0, L]$. Let $f_S \in \mathcal{F}$ be any function dependent on the sample S. Define

$$Err_D^\ell(f_S) = \mathbb{E}_{(x,y)\sim D}[\ell(f_S(x), y)],$$

and

$$Err_{m,S}^\ell(f_S) = \frac{1}{m}\sum_{i=1}^m \ell(f_S(x_i), y_i).$$

Define $S^i = \{(x_j, y_j)\}_{j\neq i} \cup (x_i', y_i')$ as a set of samples that differs in the ith sample from S, and where the differing sample is drawn from the same distribution D. We say that f_S has a *loss stability* of ε w.r.t. ℓ if $\forall (x, y) \in D$,

$$|\ell(f_S(x), y) - \ell(f_{S^i}(x), y)| \leq \varepsilon. \tag{1}$$

If f_S has a loss stability of ε w.r.t. ℓ and let $0 < \delta < 1$, then prove that the following holds with probability at least $1 - \delta$:

$$Err^{\ell}_{D}(f_S) \leq Err^{\ell}_{m,S}(f_S) + \varepsilon + (2m\varepsilon + L)\sqrt{\frac{\ln(\frac{1}{\delta})}{2m}}. \tag{2}$$

Hints:

(a) Define $\phi(S) = Err^{\ell}_{D}(f_S) - Err^{\ell}_{m,S}(f_S)$. Show that $\forall S, k, (x'_k, y'_k)$,

$$|\phi(S) - \phi(S^k)| \leq 2\epsilon + \frac{L}{m}.$$

(b) Applying bounded difference inequality using the bounded difference from hint (a). Show that with probability at least $1 - \delta$ (over $S \sim D^m$),

$$\phi(S) \leq \mathbb{E}_{S\sim D^m}[\phi(S)] + (2m\epsilon + L)\sqrt{\frac{\ln(1/\delta)}{2m}}.$$

(c) Show that

$$\mathbb{E}_{S\sim D^m}[\phi(S)] \leq \epsilon.$$

2. **(20 pts)** Letting $\mathcal{X}, \mathcal{W} \subset \mathbb{R}^d$, $\mathcal{Y} \subset \mathbb{R}$ with $\sup_{x\in\mathcal{X}} \|x\|_2 \leq R$, for some finite constant $R \geq 0$, consider the following optimization problem

$$w_S = \underset{\|w\|_2 \leq R}{\operatorname{argmin}} \left\{ \underbrace{\frac{1}{m}\sum_{i=1}^{m} \ell(w^T x_i, y_i)}_{L_S(w)} + \frac{\lambda}{2}\|w\|_2^2 \right\}, \tag{3}$$

where $\ell(\cdot, y)$ is convex and L-Lipschitz for all $y \in \mathcal{Y}$. We will show that w_S has a stability of $\frac{4LR}{\lambda m}$ (not the same as loss stability).

Define $S^i = \{(x_j, y_j)\}_{j\neq i} \cup (x'_i, y'_i)$ as a set of samples that differs in the ith sample from S, and where the differing sample is drawn from the same distribution D. Define $\bar{w} = \frac{w_{S^i} + w_S}{2}$.

(a) First establish that,

$$2L_S(\bar{w}) \leq L_S(w_S) + L_S(w_{S^i}). \tag{4}$$

(b) Using the previous part, and the optimality of w_S, w_{S^i} wrt. $L_S(w)$ and $L_{S^i}(w)$ respectively, show that

$$\frac{\lambda}{2}\|w_S\|_2^2 + \frac{\lambda}{2}\|w_{S^i}\|_2^2 - \lambda\|\bar{w}\|_2^2 \leq L_S(w_{S^i}) - L_{S^i}(w_{S^i}) + L_{S^i}(\bar{w}) - L_S(\bar{w}). \tag{5}$$

(c) Using the previous part and the Lipschitz property on the loss, show that

$$\frac{\lambda}{4}\|\Delta w\|_2^2 \leq \frac{LR}{m}\|\Delta w\|_2 \implies \|\Delta w\|_2 \leq \frac{4LR}{\lambda m}, \tag{6}$$

where $\Delta w = w_{S^i} - w_S$.

(d) Show that $f_S(\cdot) = \langle w_S, \cdot\rangle$ has a loss stability of $\frac{4L^2R^2}{\lambda m}$.

2 Lower Bound on the Oracle Complexity [40 pts]

In this problem, we will go through the proof of a lower bound on the oracle complexity of convex optimization. We will walk through some parts of the proof of Theorem 1 in [1]. The goal is to minimize a convex function $f(x)$ defined over a convex set S:

$$x_f^* \in \arg\min_{x \in S} f(x).$$

In general, it is hard to perform classical complexity analysis for convex optimization problem due to the difficulty of modeling optimization algorithms as Turing machines. An alternative is to analyze the oracle complexity of the algorithm. An *oracle* is a (possibly random) function $\phi : S \times \mathcal{F} \to \mathcal{I}$ such that for any query $x \in S$ on $f \in \mathcal{F}$, the oracle provides an answer $\phi(x, f)$ in an information set $\mathcal{I}$. Given the number of rounds T, and a function $f \in \mathcal{F}$, an optimization algorithm $\mathcal{M}$ at any step $t = 1, \ldots, T$ makes a query $x_t \in S$, and the oracle returns $\phi(x_t, f)$. The class of optimization algorithms $\mathcal{M}$ that make T queries is denoted by $\mathbb{M}_T$. We first define the following quantities:

1. Error of an algorithm $\mathcal{M} \in \mathbb{M}_T$ on function f after T steps:

$$\epsilon(\mathcal{M}, f, S, \phi) = f(x_T) - \inf_{x \in S} f(x) = f(x_T) - f(x_f^*).$$

2. The minimax error over the class of functions $\mathcal{F}$ and the class $\mathbb{M}_T$ of optimization methods taking T oracle queries:

$$\epsilon^*(\mathcal{F}, S, \phi) = \inf_{\mathcal{M}_T \in \mathbb{M}_T} \sup_{f \in \mathcal{F}} \mathbb{E}_\phi[\epsilon(\mathcal{M}, f, S, \phi)].$$

3. The worst-case error average over $\mathbb{S} = \{S \subseteq \mathbb{R}^d | S \text{ convex}, \forall x, y \in S, ||x - y||_\infty \leq 1\}$:

$$\epsilon^*(\mathcal{F}, \phi) = \sup_{S \in \mathbb{S}} \epsilon^*(\mathcal{F}, S, \phi).$$

We consider a class of oracles $\mathcal{O}$ which returns pairs of noisy functions and gradient evaluations:

$$\phi(x, f) = (\widehat{f}(x), \widehat{g}(x))$$

where $\mathbb{E}[\widehat{f}(x)] = f(x)$ and $\mathbb{E}[\widehat{g}(x))] = \nabla f(x)$. We want to show that for the class of bounded, convex, 1−Lipschitz functions $\mathcal{F}^C$ in $\mathbb{R}^d$, there exists a constant C such that

$$\sup_{\phi \in \mathcal{O}} \epsilon^*(\mathcal{F}, \phi) \geq C\sqrt{\frac{d}{T}}.$$

To obtain the lower bound, we need to a construct a subclass of function $\mathcal{G} \subseteq \mathcal{F}$. Define a distance (not necessarily a metric) ρ to be

$$\rho(f, g) = \inf_{x \in S}[f(x) + g(x) - f(x_f^*) - g(x_g^*)], \forall f \neq g \in \mathcal{F}.$$

Given $\delta \in [0, \frac{1}{4}]$, we are given a subclass of functions $\mathcal{G}(\delta) = \{g_\alpha \mid \alpha \in \mathcal{V}\}$ where $\mathcal{V} \subseteq \{-1, +1\}^d$ is used to index $\mathcal{G}(\delta)$. g_α is defined as,

$$g_\alpha(x) = \frac{1}{d} \sum_{i=1}^{d} \left(\frac{1}{2} + \alpha_i \delta\right) f_i^+(x) + \left(\frac{1}{2} - \alpha_i \delta\right) f_i^-(x),$$

where $\{f_i^+, f_i^-\}_{i=1}^d$ are some basis functions in $\mathcal{F}$. Define $\psi(\mathcal{G}(\delta)) = \min_{\alpha \neq \beta} \rho(g_\alpha, g_\beta)$ and assume $\psi(\mathcal{G}(\delta)) \geq \frac{\delta}{2}, |\mathcal{V}| \geq (2/\sqrt{e})^{d/2}$. We encourage you to see [1] for an exact construction of the basis functions, although it is not necessary to solve this problem. We can now restrict our analysis to the subset $\mathcal{G}(\delta) \subseteq \mathcal{F}$, because the minimax error over a larger set can only be higher. Since we want to prove a lower bound of $\epsilon^*(\mathcal{F}, \phi)$ which is a sup over the set of first order oracles $\phi \in \mathcal{O}$, it is sufficient to prove the lower bound for a specific oracle. Consider the oracle ϕ that presents noisy value and gradient samples from g_α, according to the following process:

Given a function $g_\alpha \in \mathcal{G}(\delta)$ unknown to the optimization algorithm,

- Pick an index $i_t \in \{1, ..., d\}$ uniformly at random.
- Draw b_{i_t} from a Bernoulli distribution with parameter $\frac{1}{2} + \alpha_{i_t}\delta$.
- Return the value and subgradient of $\widehat{g}_\alpha(x) = b_{i_t} f_{i_t}^+ + (1 - b_{i_t}) f_{i_t}^-$.

We can see that $\phi \in \mathcal{O}$ since $\mathbb{E}[\widehat{g}_\alpha(x)] = g_\alpha(x)$ and $\mathbb{E}[\nabla \widehat{g}_\alpha(x)] = \nabla g_\alpha(x)$.

Let $\mathcal{M}_T$ be any algorithm making at most T queries achieving the minimax error. Informally, the proof contains three steps: (i) $\mathcal{M}_T$ obtaining a low minimax optimization error ϵ^* over $\mathcal{G}(\delta)$ implies that the true $\alpha^* \in \mathcal{V}$ can be identified using $\mathcal{M}_T$; (ii) identifying $\alpha^* \in \mathcal{V}$ can be related to a hypothesis testing problem on Bernoulli random variables (coin toss problem); (iii) Using generalized Fano's inequality to lower bound the minimax error for hypothesis testing.

1. **(8 pts)** Show that for any $x \in S$, there can be at most one function $g_\alpha \in \mathcal{G}(\delta)$ for which $g_\alpha(x) - g_\alpha(x^*_{g_\alpha}) \leq \frac{\psi(\mathcal{G}(\delta))}{3}$.

2. **(8 pts)** Suppose an algorithm $\mathcal{M}_T$ achieves a minimax optimization error upper bounded as

$$\epsilon^*(\mathcal{M}_T, \mathcal{G}(\delta), S, \phi) = \sup_{f \in \mathcal{G}(\delta)} \mathbb{E}[\epsilon(\mathcal{M}_T, f, S, \phi)] \leq \frac{\psi(\mathcal{G}(\delta))}{9},$$

   then show that such a method $\mathcal{M}_T$ can be used to construct an estimator $\widehat{\alpha}(\mathcal{M}_T)$ such that

$$\max_{\alpha^* \in \mathcal{V}} \mathbb{P}_\phi[\widehat{\alpha} \neq \alpha^*] \leq \frac{1}{3}.$$

   Hence we have reduced the problem of achieving a small minimax optimization error to a hypothesis test, which we will bound using a version of Fano's inequality. We will choose δ based on the desired error ϵ such that is satisfies $\psi(\mathcal{G}(\delta)) \geq 9\epsilon$.

3. **(14 pts)** We now lower bound $\mathbb{P}_\phi[\widehat{\alpha} \neq \alpha^*]$. The problem of identifying the true α^* can be related to identifying a Bernoulli distribution using T samples. The first order oracle ϕ described earlier performs a coin toss and returns a function value and a subgradient based on the coin toss. Providing the coin toss outcome b_{i_t} makes the problem of identifying α^* no harder than providing function values and the subgradient of $\widehat{g}(\alpha)$ since the function values and the subgradients can be computed using the coin toss outcome itself.

   We now consider a different oracle that instead returns the coin toss outcomes (i, b) where $i \in \{1, \cdots, d\}$ denotes the randomly chosen index and $b \in \{0, 1\}$ denotes the outcome of the coin toss. For any $\alpha^* \in \mathcal{V}$, consider $\theta^* = (\frac{1}{2} + \alpha_1^*\delta, \ldots, \frac{1}{2} + \alpha_d^*\delta)$. Suppose for a total of T times, we toss a set of d coins with biases given by θ^*. At each step the outcome of only one coin chosen uniformly at random is revealed by the oracle as a pair (i, b). The goal of the coin toss problem is to identify θ^* using these random samples. The coin toss problem being an easier problem, is sufficient to compute minimax lower bounds for this.

   Let $\widehat{\theta}$ be an estimator for the coin toss problem making at most T queries. Then, as argued above $\inf_{\widehat{\theta}} \max_{\theta^*} \mathbb{P}[\widehat{\theta} \neq \theta^*] \leq \inf_{\widehat{\alpha}} \max_{\alpha^*} \mathbb{P}[\widehat{\alpha} \neq \alpha^*]$. Show that for all $\delta \leq 1/4$,

$$\inf_{\widehat{\alpha}} \max_{\alpha^* \in \mathcal{V}} \mathbb{P}[\widehat{\alpha} \neq \alpha^*] \geq \left\{ 1 - \frac{16T\delta^2 + \log 2}{\frac{d}{2}\log(2/\sqrt{e})} \right\}$$

4. **(10 pts)** Now set $\delta = 18\epsilon$ (recall that we should set δ based on the desired error ϵ), and use the obtained upper and lower bound of $\mathbb{P}(\widehat{\alpha} \neq \alpha^*)$ to get that for all $d \geq 11$, $\epsilon \geq c\sqrt{\frac{d}{T}}$ for some constant c. Hence,

$$\sup_{\phi \in \mathcal{O}} \epsilon^*(\mathcal{F}, \phi) \geq \Omega\left(\sqrt{\frac{d}{T}}\right).$$

References

[1] Alekh Agarwal, Martin J Wainwright, Peter L Bartlett, and Pradeep K Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*, pages 1–9, 2009.