CSC 588: Homework 4

Chicheng Zhang

April 8, 2022

Please complete the following exercises and read the following instructions carefully.

- Your solutions to these problems will be graded based on both correctness and clarity. Your arguments should be clear: there should be no room for interpretation about what you are writing. Otherwise, I will assume that they are wrong, and grade accordingly.
- If you feel unable to make progress on any of the questions, you can post your questions on Piazza. Try posing your questions to be as general as possible, so that it can promote discussion among the class.
- You are encouraged to discuss the homework questions with your classmates, but the discussions should only be at a high level, and you should write your solutions in your own words. For every question you have had discussions on, please mention explicitly whom you have discussed with; otherwise it may be counted as academic integrity violation.
- For detailed homework policies, please read the course syllabus, available on the course website.

This homework is due on Apr 26, 2022, 5pm MST, on gradescope.

Problem 1

Solve the following in-class exercises:

1. In the class, we have seen that, if distribution D is supported on $\mathcal{X} = \left\{x \in \mathbb{R}^d : \|x\|_\infty \leq X_\infty\right\}$, and consider linear hypothesis class $\mathcal{W} = \left\{w \in \mathbb{R}^d : \|w\|_1 \leq W_\infty\right\}$, ℓ_1/ℓ_∞ margin bound is a factor of $\sqrt{d}$ tighter than ℓ_2/ℓ_2 margin bound.

   Can you find another distribution supported on another domain $\mathcal{X}'$ and another hypothesis class $\mathcal{W}'$ of linear classifiers, such that when applied to them, ℓ_2/ℓ_2 margin bound is a factor of $\sqrt{d}$ tighter than ℓ_1/ℓ_∞ margin bound instead? Justify your answer.

2. Some exercises on convexity:

   (a) Prove: If f is a convex function with domain $\mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$ is a matrix, and $b \in \mathbb{R}^m$ is a vector, then $g(x) := f(Ax + b)$ is also a convex function.

   (b) Prove: If functions $f_1, \ldots, f_n$ are convex with domains $\mathbb{R}^m$, then $h(x) := \max_{i=1}^n f_i(x)$ is also convex.

   (c) Prove: if f is a λ-strongly convex function with respect to $\|\cdot\|_2$ on convex domain Ω ($\lambda > 0$), then its has a unique minimizer – that is, for any $u, v \in \arg\min_{w \in \Omega} f(w)$, it must be the case that $u = v$.

   (d) Consider convex function $f(x) = x$ with domain $\Omega = [0, 1]$. What is $\partial f(x)$? Now consider its minimizer $x^* = \arg\min_{x \in \Omega} f(x)$ – is it true that $0 \in \partial f(x^*)$?

Problem 2

Show that for AdaBoost, at iteration t, the updated distribution D_{t+1} satisfies that

$$\sum_{i=1}^{m} D_{t+1}(i) I(h_t(x_i) \neq y_i) = \frac{1}{2}.$$

Intuitively, why is this formula reasonable?

Problem 3

In the class we have seen that ℓ_2-regularization can induce stability and good generalization performance. This exercise explores stability properties of entropy regularization with different geometry in data. Consider a logistic regression setting, where we have a distribution D over $\mathcal{X} \times \mathcal{Y}$, where the feature space $\mathcal{X} = \left\{x \in \mathbb{R}^d : \|x\|_\infty \leq R_\infty\right\}$ and the label space $\mathcal{Y} = \{\pm 1\}$. Let the hypothesis set

$$\mathcal{H} = \Delta^{d-1} = \left\{ w \in \mathbb{R}^d : \forall j, w_j \geq 0, \sum_{j=1}^{d} w_j = 1 \right\}.$$

Consider the logistic loss $\ell(w, (x, y)) = \ln\Big(1 + \exp\big(-y \langle w, x\rangle\big)\Big)$, and denote by $L_D(w) = \mathbb{E}_{(x,y)\sim D}\left[\ell(w, (x, y))\right]$ the generalization loss of w. A set of training examples $S = ((x_1, y_1), \ldots, (x_m, y_m))$ is drawn iid from D, and the learning algorithm $\mathcal{A}(\lambda)$ is defined as:

Given input S, return

$$\hat{w} = \arg\min_{w\in\mathcal{H}} \left(F_S(w) := \lambda\psi(w) + \frac{1}{m}\sum_{i=1}^{m} \ell(w, (x_i, y_i)) \right),$$

where $\psi(w) = \sum_{j=1}^{d} w_j \ln w_j$ is the negative entropy regularizer.

Answer the following questions:

1. Show that:
   (a) For every $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and any w_1, w_2, $|\ell(w_1, (x, y)) - \ell(w_2, (x, y))| \leq R\|w_1 - w_2\|_1$; in other words, $\ell(w, (x, y))$ is R-Lipschitz with respect to w and $\|\cdot\|_1$.

   (b) For any vector $w \in \left\{w \in \mathcal{H} : \forall j, w_j > 0\right\}$ and any vector $x \in \mathbb{R}^d$, $x^\top \nabla^2\psi(w)x \geq \|x\|_1^2$; here $\nabla^2 G(w) \in \mathbb{R}^{d\times d}$ denotes the Hessian of function G at w.

   (c) For all $w \in \mathcal{H}$, $\psi(w) \in [-\ln d, 0]$.

2. Show that $\mathcal{A}(\lambda)$ is $\frac{2R^2}{\lambda m}$-OARO stable. You may want to use first order optimality condition for convex optimization and second order Taylor expansion[1] to solve this problem. You can also assume without proof that for any training sample S, the corresponding regularized ERM $\hat{w}$ satisfies $\hat{w}_j > 0$ for all j.

3. Finally, provide an upper bound on $\mathbb{E}\left[L_D(\hat{w})\right] - \min_{w'\in\mathcal{H}} L_D(w')$. How would you choose λ to minimize your bound?

[1] see e.g. Theorem 5 of `http://www.math.toronto.edu/courses/mat237y1/20199/notes/Chapter2/S2.6.html`

Problem 4

In this exercise, we conduct an empirical analysis on the effect of step size in online (stochastic) gradient descent (abbrev. OGD). Please submit your source code by emailing to `csc588homeworks@gmail.com`. Some preparations:

1. For $d \in \mathbb{N}_+$, let $u = \frac{2}{\sqrt{d}}(1, 1, \ldots, 1) \in \mathbb{R}^d$. Define the following distribution D over binary classification examples: x is drawn uniformly from $[-1, +1]^d$; given x, $\mathbb{P}(Y = y \mid X = x) = \frac{1}{1+\exp\left(-y\langle u,x\rangle\right)}$ for $y \in \{-1, +1\}$. Write a program that draws random examples from D.

2. Recall the logistic loss $\ell(w, (x, y)) = \ln\Big(1 + \exp\big(-y \langle w, x\rangle\big)\Big)$. Write a program that takes input step size $\eta > 0$, and runs OGD on one pass of the logistic losses induced by training examples, with constraint set $\Omega = \left\{w \in \mathbb{R}^d : \|w\|_2 \leq 20\right\}$ and initializer $w_1 = \vec{0}$. That is, it runs OGD with $\left\{f_t(w) = \ell(w, (x_t, y_t))\right\}_{t=1}^T$, where (x_t, y_t) is the t-th training example. The program should return the average iterate $\hat{w} = \frac{1}{T}\sum_{t=1}^T w_t$.

Denote by $L_D(w) = \mathbb{E}_{(x,y)\sim D}\ell(w, (x, y))$. Answer the following questions:

1. Given any d, can you provide a theoretical upper bound on $\mathbb{E}\left[L_D(\hat{w})\right] - \min_{w' \in \Omega} L_D(w')$, when step size η is used? How would you choose the "theoretically-best" η based on this upper bound?

2. Fix $d = 10$. For every $c \in C = \{0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, 100\}$, repeat the following process 5 times:

   (a) Draw $T = 100$ training and $T_{\text{test}} = 500$ test examples iid from D.

   (b) Run the above one-pass SGD program with $\eta = \frac{c}{\sqrt{T}}$ over the training examples, return $\hat{w}$.

   (c) Approximately evaluate $L_D(\hat{w})$ using empirical average over the test examples.

   Calculate the mean and the standard deviation of $L_D(\hat{w})$ over the 5 runs. Plot these values as a function of c, using log scale on the x-axis, and set y axis limit to $[0, 2]$ (you can use the error bar or "fill-between" functionalities provided by many plotting libraries). For reference, also plot a horizontal lines for $L_D(u)$ - this is the minimum achievable logistic loss on D. Is the c that minimizes $L_D(\hat{w})$ (within C) comparable to the theoretically-optimal value?

3. Repeat the same experiments in item 2 and plot the same graphs, this time for $d = 100$ and $d = 1000$. How do the optimal choices of c in these experiments compare with the previous experiment?

Problem 5

How much time did it take you to complete this homework?