CSC 588: Homework 3

Chicheng Zhang

March 29, 2021

- This homework is due on Apr 15 on gradescope.
- 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.
- Feel free to use existing theorems from the course notes / the textbook.

Problem 1

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 $\mathcal{X} = \left\{x \in \mathbb{R}^d : \|x\|_\infty \leq R\right\}$ 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 $\hat{w} = \arg\min_{w\in\mathcal{H}}\left(F_S(w) := \lambda\psi(w) + \sum_{i=1}^m \ell(w, (x_i, y_i))\right)$, 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, $|\ell(w, (x, y)) - \ell(w_2, (x, y))| \leq R\|w_1 - w_2\|_1$; in other words, $\ell(w, (x, y))$ is R-Lipschitz with respect to w and $\|\cdot\|_1$.
   (b) For any vector $w \in \mathcal{H}$ 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 to solve this problem.
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?

Problem 2

In this problem, we provide a refined analysis of the Perceptron algorithm from an online gradient descent perspective. Consider the following variant of Perceptron algorithm (call it Perceptron(η)):

Step size $\eta > 0$.
Initialize $w_1 \leftarrow \vec{0} \in \mathbb{R}^d$.
for $t = 1, 2, \ldots, T$ **do**
  Receive example $x_t \in \mathbb{R}^d$
  Predict $\hat{y}_t = \text{sign}(\langle w_t, x_t \rangle)$
  Receive label $y_t \in \{-1, +1\}$
  if $\hat{y}_t \neq y_t$ **then**
    $M_t \leftarrow 1$, $w_{t+1} \leftarrow w_t + \eta y_t x_t$
  else
    $M_t \leftarrow 0$, $w_{t+1} \leftarrow w_t$
  end if
end for

In addition, suppose the learner is faced with an oblivious adversary, that is, $(x_t, y_t)_{t=1}^T$ are chosen before the interaction starts; also, for all t, $\|x_t\|_2 \leq R$. Answer the following questions:

1. Show that different choices of step size $\eta > 0$ won't affect the value of $\sum_{t=1}^T M_t$, the total number of mistakes made by the algorithm.

2. Define $f_t(w) = M_t \langle w, -y_t x_t \rangle$. Suppose the learner runs Perceptron(η) with $\eta > 0$; provide an upper bound of $R_T(w^\star) = \sum_{t=1}^T f_t(w_t) - \sum_{t=1}^T f_t(w^\star)$ for any $w^\star$.

3. Now suppose the learner runs Perceptron($\eta = 1$). Show that for any $w^\star \in \mathbb{R}^d$,

$$\sum_{t=1}^T M_t \leq \sum_{t=1}^T M_t \ell(w^\star, (x_t, y_t)) + \sqrt{\|w^\star\|^2 R^2 \left(\sum_{t=1}^T M_t \right)},$$

   where $\ell(w, (x, y)) = \max(0, 1 - y \langle w, x \rangle)$ is the hinge loss of w on (x, y). Conclude that

$$\sum_{t=1}^T M_t \leq \min_{w^\star \in \mathbb{R}^d} \left(\sum_{t=1}^T \ell(w^\star, (x_t, y_t)) + \sqrt{\sum_{t=1}^T \ell(w^\star, (x_t, y_t)) \cdot \|w^\star\|_2^2 R^2} + \|w^\star\|_2^2 R^2 \right).$$

4. Additionally, suppose there exists some u such that $\|u\| = 1$, and for all t, $y_t \langle u, x_t \rangle \geq \gamma$. Show that Perceptron($\eta = 1$) guarantees that $\sum_{t=1}^T M_t \leq \frac{R^2}{\gamma^2}$.

Problem 3

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 `chichengz@cs.arizona.edu`. 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 $\Omega = \big\{w \in \mathbb{R}^d : \|w\|_2 \le 20\big\}$ and initializer $w_1 = \vec{0}$. That is, it runs OGD with $\big\{f_t(w) = \ell(w,(x_t,y_t))\big\}_{t=1}^{T}$, 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 $\bar{w}_T$.

   (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 $L_D(u)$ - these are the minimum achievable logistic loss and 0-1 error 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 4

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