CS7545, Fall 2018: Machine Learning Theory - Homework #3

Jacob Abernethy, Bhuvesh Kumar, and Jun-Kun Wang Due: Wednesday, October 17, 2018 at 2 p.m

Homework Policy: Working in groups is fine, but *every student* must submit their own writeup. Please write the members of your group on your solutions. There is no strict limit to the size of the group but we may find it a bit suspicious if there are more than 4 to a team. Questions labelled with **(Challenge)** are not strictly required, but you'll get some participation credit if you have something interesting to add, even if it's only a partial answer.

1) **Generalized Minimax Theorem.** Let $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ be convex compact sets. Let $f : X \times Y \to \mathbb{R}$ be some differentiable function with bounded gradients, where $f(\cdot, \mathbf{y})$ is convex in its first argument for all fixed $\mathbf{y}$, and $f(\mathbf{x}, \cdot)$ is concave in its second argument for all fixed $\mathbf{x}$.

Prove that

$$\inf_{\mathbf{x}\in X} \sup_{\mathbf{y}\in Y} f(\mathbf{x}, \mathbf{y}) = \sup_{\mathbf{y}\in Y} \inf_{\mathbf{x}\in X} f(\mathbf{x}, \mathbf{y}).$$

Furthermore, give an efficient algorithm for finding an ϵ-optimal pair $(\mathbf{x}^*, \mathbf{y}^*)$ for any parameter $\epsilon > 0$.

2) **Regret of Follow the Perturbed Leader.** We will observe a sequence of loss vectors $\boldsymbol{\ell}^1, \boldsymbol{\ell}^2, \ldots, \boldsymbol{\ell}^T \in [0, 1]^n$. We need an algorithm for picking a sequence of distributions $\mathbf{p}^1, \mathbf{p}^2, \ldots, \mathbf{p}^T \in \Delta_n$ with the goal of minimizing regret. For the rest of this problem we shall define regret relative to some $\mathbf{p}$ as

$$\text{Regret}_T(\text{Alg}; \mathbf{p}) := \sum_{t=1}^{T} (\mathbf{p}^t \cdot \boldsymbol{\ell}^t - \mathbf{p} \cdot \boldsymbol{\ell}^t).$$

Note that this differs slightly than our usual notion where $\mathbf{p}$ is chosen to be the best distribution (or expert) in hindsight.

I have already mentioned an algorithm often called Follow The Leader (FTL) defined as

$$\text{FTL} \quad := \quad \mathbf{p}^t \leftarrow \arg\min_{\mathbf{p}\in\Delta_n} \mathbf{p} \cdot \left(\sum_{s=1}^{t-1} \boldsymbol{\ell}^s \right)$$

There's an easy lower bound that shows that this algorithm can achieve $\Theta(T)$ regret which is bad! But what if we just perturb this algorithm slightly? Here's an alternative approach which involves playing FTL on the cumulative loss vector with some added noise.

$$\text{FTPL} \quad := \quad X \overset{\text{u.a.r.}}{\sim} [0, b]^n; \qquad \text{then } \forall t \quad \mathbf{p}^t \leftarrow \arg\min_{\mathbf{p}\in\Delta_n} \mathbf{p} \cdot \left(X + \sum_{s=1}^{t-1} \boldsymbol{\ell}^s \right)$$

Note that the perturbation X is only sampled *once* in this algorithm. X is sampled uniformly at random from a cube, and note that the sidelength of the cube $b > 0$ is a parameter which we can tune.

For analysis purposes, it is convenient to define two *fictitious* algorithms, known as Be The Leader (BTL), and Be The Perturbed Leader (BTPL).

$$\begin{aligned} \text{BTL} &:= \mathbf{p}^t \leftarrow \arg\min_{\mathbf{p}\in\Delta_n} \mathbf{p}\cdot\left(\sum_{s=1}^{t} \boldsymbol{\ell}^s\right) \\ \text{BTPL} &:= X \overset{\text{u.a.r.}}{\sim} [0,b]^n; \quad \text{then } \forall t \quad \mathbf{p}^t \leftarrow \arg\min_{\mathbf{p}\in\Delta_n} \mathbf{p}\cdot\left(X + \sum_{s=1}^{t} \boldsymbol{\ell}^s\right). \end{aligned}$$

What is different here? Notice I changed the sum to end at $s = t$ rather than $s = t - 1$ – that's why these algorithms are fictitious, they get to see one datapoint in the future! Here we are "being" the leader rather than "following" the leader because we actually can compute the leader up to *and including* the loss vector that will arrive today.

(a) BTL, while not a realistic algorithm, kicks ass! Prove, for any $\mathbf{p} \in \Delta_n$, that

$$\text{Regret}_T(\text{BTL}; \mathbf{p}) \leq 0$$

Hint: Induction.

(b) BTPL is really not that much worse than BTL. Prove that

$$\text{Regret}_T(\text{BTPL}; \mathbf{p}) \leq b.$$

Note that this is a deterministic statement, doesn't depend on the sample of X.

(c) Assume that BTPL and FTPL were run using the same perturbation X sampled before round 1. Let $\mathbf{p}^1, \mathbf{p}^2, \ldots, \mathbf{p}^T$ be the distributions played by FTPL throughout the sequence. Prove that for any perturbation X,

$$\text{Regret}_T(\text{FTPL}; \mathbf{p}) = \text{Regret}_T(\text{BTPL}; \mathbf{p}) + \sum_{t=1}^{T} (\mathbf{p}^t - \mathbf{p}^{t+1}) \cdot \boldsymbol{\ell}^t.$$

Again this is for a fixed, arbitrary X.

(d) It turns out that by perturbing the loss by X, we are much less likely to switch from round to round. If $\mathbf{p}^t, \mathbf{p}^{t+1}$ are the distributions played by FTPL on rounds t and $t+1$ (respectively), then show that for any t,

$$\mathop{\mathbb{E}}_{X \overset{\text{u.a.r.}}{\sim} [0,b]^n} [(\mathbf{p}^t - \mathbf{p}^{t+1}) \cdot \boldsymbol{\ell}^t] \leq \frac{n}{b}.$$

Hint: Define the random variables $Z_t := X + \sum_{s=1}^{t-1} \boldsymbol{\ell}^s$ and $Z_{t+1} := X + \sum_{s=1}^{t} \boldsymbol{\ell}^s$. Notice that the distributions of Z_t and Z_{t+1} overlap significantly.

(e) Let's put it all together! Prove that for a particular choice of b we can achieve:

$$\mathop{\mathbb{E}}_{X \overset{\text{u.a.r.}}{\sim} [0,b]^n} [\text{Regret}(\text{FTPL}; \mathbf{p})] \leq \sqrt{nT}.$$

It's ok if you didn't solve all of the above, you may use the conclusions from each subproblem.

(f) **(Challenge)** It is important for the analysis that X is sampled once and fixed throughout the sequence. But in terms of expected regret, would it matter if we sampled X separately for each round? Why or why not?

(g) (**Challenge**) It's too bad the above bound isn't as tight as the $O(\sqrt{T \log n})$ bound we can get with EWA. Can FTPL be improved using a better choice of perturbation X? I might suggest a Laplace distribution or a Gaussian.

(h) (**Challenge**) Is there a way to implement EWA (exponential weights algorithm) in the action setting (i.e. the "hedge" setting) using FTPL? In other words, can you choose a perturbation random variable X such that the maximizing action on round t is chosen with the same probabilities as the EWA distribution?

(i) (**Challenge**) Show that any FTPL can be formulated as FTRL, although the regularizer for FTRL may not be efficiently computable. *Hint*: Given FTPL with distribution $\mathcal{D}$, consider the Fenchel conjugate of the function

$$g(\mathbf{L}) := \mathbb{E}_{X \sim \mathcal{D}} \left[\max_{\mathbf{p} \in \Delta_n} -\mathbf{p} \cdot (X + \mathbf{L}) \right].$$

3) **Online Non-Convex Optimization.** Sometimes our nice assumptions don't always hold. AWWW SHUCKS!! But maybe things will still work out just fine. For the rest of this problem assume that $X \subset \mathbb{R}^n$ is the learner's decision set, and the learner observes a sequence of functions $f_1, f_2, \ldots, f_T$ mapping $X \to \mathbb{R}$. The regret of an algorithm choosing a sequence of $\mathbf{x}_1, \mathbf{x}_2, \ldots$ is defined in the usual way:

$$\text{Regret}_T := \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x} \in X} \sum_{t=1}^{T} f_t(\mathbf{x})$$

Wouldn't it ruin your lovely day if the functions f_t were not convex? Maybe the only two conditions you can guarantee is that the functions f_t are bounded (say in $[0,1]$) and are 1-*Lipschitz*: they satisfy that $|f_t(\mathbf{x}) - f_t(\mathbf{x}')| \leq \|\mathbf{x} - \mathbf{x}'\|_2$. Prove that, assuming X is convex and bounded, there exists a randomized algorithm with a reasonable expected-regret bound. Something like $\mathbb{E}[\text{Regret}_T] \leq \sqrt{nT \log T}$ would be admirable. (Hint: Always good to ask the *experts* for ideas. And you needn't worry about efficiency.)

4) **Perceptron extension for GLM.** A generalization of the linear regression problem is the Generalized Linear Modelling (GLM) problem where we are presented with labeled data $(\mathbf{x}_i, y_i)$ where y_i is generated as $y_i = g(\langle \mathbf{w}^*, \mathbf{x}_i \rangle)$ where $g : \mathbb{R} \to \mathbb{R}$ is the inverse link function and $\mathbf{w}^*$ is the underlying model. We will analyze a percepton style algorithm for GLM call GLMtron. Note that g is known to the algorithm.

Input: Inverse link fucntion g, step length η
$\mathrm{w}^t \leftarrow \mathbf{0}$
for $t = 1, \cdots, T$ **do**
  Receive data point $\mathbf{x}^t$
  Predict $\hat{y}^t = g(\langle \mathbf{w}^t, \mathbf{x}_i \rangle)$
  Receive true label $y^t \in \mathbb{R}$
  Incur loss $\ell^t = (\hat{y}^t - y^t)^2$
  Update $\mathbf{w}^{t+1} = \mathbf{w}^t + \eta\,(y^t - \hat{y}^t)\,\mathbf{x}^t$
end

Algorithm 1: GLMtron

The total error incurred by GLMtron in T time steps will be denoted by M_T, i.e

$$M_T := \sum_{t=1}^{T} \ell^t = \sum_{t=1}^{T} \left(\hat{y}^t - y^t\right)^2$$

(a) Show that if $g(a) = sign(a)$ and $y^t \in \{-1, +1\}$, then GLMtron reduces to perceptron for a suitable choice of η.

For the following parts, you may assume $\|\mathbf{x}^t\|_2 \leq 1$ for all t. Also assume g is a non decreasing function and that g is an invertible function that satisfies $K\,|a-b| \leq |g(a)-g(b)| \leq L\,|a-b|$ for all $a, b \in \mathbb{R}$.

(b) Show that if the problem is realizable, i.e there exists some $\mathbf{w}^*$ such that $y^t = g(\langle \mathbf{w}^*, \mathbf{x}^t \rangle)$ for all t, then M_T is bounded for an appropriate choice of η. What is the bound? Does it depend on T and η? You may assume $\|\mathbf{w}^*\|_2 \leq 1$. **Hint:** It may be useful to prove that for all $a, b \in \mathbb{R}$, $(a-b) \cdot (g(a) - g(b)) \geq \frac{1}{L}\left(g(a) - g(b)\right)^2$

(c) (**Challenge**) Perform the analysis in a non-realizable setting. It might be useful to compare M_T with the mistake of a fixed benchmark predictor $\mathbf{w}^*$ which has the minimum mistake M_T^*. Note that you cannot assume that $\|\mathbf{w}^*\|_2 \leq 1$.