CS591 E2: Optimization Methods and their Applications – Fall 2016

Homework 3

Due date: November 17, 2016

Homework Policy and Guidelines

You are encouraged to collaborate on the solution of the homeworks and to consult any materials, but you must write up your own answers and you must acknowledge all of your collaborators and sources.

The problems marked with a $(*)$ are more challenging. You may not be able to completely solve some of the more challenging problems, that is completely normal!

Some of the problems ask you to fill in a proof that we did not cover in class; the readings will often have these proofs, you are free to consult them but you must write up a complete proof in your own words. In general it may be good to keep in mind that some of the proofs in the textbooks may leave some steps to the reader, and it is very important to make sure that you know how to fill in those missing steps. Also, thinking about those proofs on your own will help you understand the material better.

Problem 1. Recall our toy stock market example. Prove that, in the worst case, no *deterministic* algorithm can make less than $2M^*(T) + \lfloor \log_2 n \rfloor$ mistakes, where n is the number of experts and $M^*(T)$ is the number of mistakes made by the best expert over T rounds.

Problem 2. Recall the learning with expert advice framework. Let n be the total number of experts. Show that, for any (non-empty) subset U of the experts, the multiplicative weights update method suffers a total loss $L(T)$ of at most

$$L(T) \leq \max_{i \in U} \left(\sum_{t=1}^{T} \ell_i^t + \eta \sum_{t=1}^{T} |\ell_i^t| \right) + \frac{\ln\left(\frac{n}{|U|}\right)}{\eta},$$

where $\ell_i^t \in [-1, 1]$ is the loss of expert i in round t and $0 < \eta \leq 1/2$ is the step size of the multiplicative weights update method.

Note: this means that the multiplicative weights update method has better performance when there are many good experts.

Problem 3 (Online Gradient Descent revisited)**.** In this problem, we revisit the Online Gradient Descent algorithm (OGD). In this problem, we will consider a prediction domain $S \subseteq \mathbb{R}^d$ (as usual, S is a convex set). In class, we used a fixed step size η in the OGD algorithm. In this problem, we allow for varying step sizes η_t. More precisely, we consider the following algorithm.

Let $f_1, f_2, \ldots, f_T$ be a sequence of loss functions. Here we assume that each f_t is σ_t-strongly-convex with respect to the ℓ_2-norm. That is, for all $w \in S$ and any subgradient $z_t \in \partial f_t(w)$, the following holds for all $u \in S$:

$$f_t(u) \geq f_t(w) + \langle z_t, u - w \rangle + \frac{\sigma_t}{2} \|u - w\|_2^2.$$

The OGD algorithm starts with an initial point $w_1 = 0 \in S$ and constructs its predictions according to the following update rule: $w_{t+1} = \Pi_S(w_t - \eta_t z_t)$ where $z_t \in \partial f_t(w_t)$. As before, Π_S is the projection onto S.

In the following, we analyze the algorithm and see how to choose the step sizes η_t. Throughout, we let u be a fixed point in S.

(a) Prove the following regret bound:

$$\text{Regret}_T(u) \leq \sum_{i=1}^{T} \left(\frac{1}{2\eta_t} \left(\|w_t - u\|_2^2 - \|w_{t+1} - u\|_2^2 \right) + \frac{\eta_t}{2} \|z_t\|_2^2 - \frac{\sigma_t}{2} \|w_t - u\|_2^2 \right).$$

(b) Now assume that we have the following bounds on the norms of the vectors in S and the subgradients: $\|u\|_2 \le B$ for all $u \in S$ and $\|z_t\| \le L$ for all t. Additionally, assume that we pick the step sizes so that $\frac{1}{\eta_t} \ge \frac{1}{\eta_{t-1}} + \sigma_t$ (we use the convention $1/\eta_0 = 0$). Use the regret bound from part (a) to show that

$$\text{Regret}_T(u) \le 2B^2 \left(\frac{1}{\eta_T} - \sum_{t=1}^{T} \sigma_t \right) + \frac{L^2}{2} \sum_{t=1}^{T} \eta_t.$$

(c) Let us use the regret bound from (b) in the following simpler setting. Let σ be such that $\sigma_t \ge \sigma > 0$ for all t. Show that, if we set the step sizes to $\eta_t = \frac{1}{\sum_{i=1}^{t} \sigma_i}$, we obtain the following bound

$$\text{Regret}_T(u) \le \frac{L^2(\log(T)+1)}{2\sigma}.$$

Note: this shows *logarithmic* regret for all strongly convex loss functions, not just quadratics, provided that the step sizes decay as $O(1/t)$. Note that this bound deteriorates as the strong convexity deteriorates.

(d) If we do not have strong convexity, we can still set the step sizes to get a meaningful regret bound. Suppose that $\sigma_t = 0$. Show that, if we set $\eta_t = \frac{B}{L\sqrt{t}}$, we obtain

$$\text{Regret}_T(u) \le 3BL\sqrt{T}.$$

Note: this bound shows that we can get a $O(\sqrt{T})$ regret even without knowing T in advance.

(e) Let us now use the analysis above to guide our choice of step sizes in a concrete setting. Suppose that at each step t, we are given a feature vector $x_t \in \mathbb{R}^d$; there is also a label $y_t \in \mathbb{R}$ that is not known to us at the beginning of round t, and our task is to predict a vector $w_t \in \mathbb{R}^d$ such that $\langle w_t, x_t \rangle$ is "close" to the true label y_t. Suppose that $\|u\|_2 \le B$ for all $u \in S$, $\|x_t\| \le C$, and $\|y_t\| \le a$. Consider the following two common loss functions:

- Linear least-squares regression: $f_t(w) = (y_t - \langle w, x_t \rangle)^2$.
- Regularized SVM: $f_t(w) = \max\{0, 1 - y_t \cdot \langle w, x_t \rangle\} + \lambda \|w\|_2^2$.

Show how to set the step sizes in each of the two scenarios. Write down η_t as a function of a, B, C, t and upper bound the regret using the analysis above. Please justify your answers.

Problem 4. Show that the entropy regularizer $R(w) = \sum_{i=1}^{d} w_i \log(w_i)$ is $\frac{1}{B}$-strongly-convex with respect to the ℓ_1-norm over the set $S = \{w \in \mathbb{R}^d \colon w > 0, \|w\|_1 \le B\}$.

Problem 5. Recall the analysis we saw in class for the Follow-the-Regularized-Leader algorithm with a strongly convex regularizer. Use this analysis to show the following guarantee for the Online Mirror Descent algorithm.

Let R be a $\frac{1}{\eta}$-strongly-convex function over S with respect to a norm $\|\cdot\|$. Suppose we run the Online Mirror Descent algorithm with a link function

$$g(x) = \text{argmax}_{w \in S} \left(\langle w, x \rangle - R(w) \right).$$

Show that, for all $u \in S$,

$$\text{Regret}_T(u) \le R(u) - \min_{v \in S} R(v) + \eta \sum_{t=1}^{T} \|z_t\|_*^2$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$: $\|x\|_* = \sup_{y \colon \|y\| \le 1} \langle y, x \rangle$. Show that, if we additionally assume that each f_t is L_t-Lipschitz, we can upper bound $\|z_t\|_* \le L_t$.