COS 511: Theoretical Machine Learning

Homework #6 Due:
Lagrange and online learning April 11, 2013

Problem 1

a. [10] In class, we argued that if a function L satisfies the "minmax property"

$$\min_{\mathbf{w}} \max_{\boldsymbol{\alpha}} L(\mathbf{w}, \boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}} \min_{\mathbf{w}} L(\mathbf{w}, \boldsymbol{\alpha}), \tag{1}$$

and if $(\mathbf{w}^*, \boldsymbol{\alpha}^*)$ are the desired solutions

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \max_{\boldsymbol{\alpha}} L(\mathbf{w}, \boldsymbol{\alpha}) \tag{2}$$

$$\boldsymbol{\alpha}^* = \arg\max_{\boldsymbol{\alpha}} \min_{\mathbf{w}} L(\mathbf{w}, \boldsymbol{\alpha}), \tag{3}$$

then $(\mathbf{w}^*, \boldsymbol{\alpha}^*)$ is a saddle point:

$$L(\mathbf{w}^*, \boldsymbol{\alpha}^*) = \max_{\boldsymbol{\alpha}} L(\mathbf{w}^*, \boldsymbol{\alpha}) = \min_{\mathbf{w}} L(\mathbf{w}, \boldsymbol{\alpha}^*). \tag{4}$$

(Here, it is understood that $\mathbf{w}$ and $\boldsymbol{\alpha}$ may belong to a restricted space (e.g., $\boldsymbol{\alpha} \geq 0$) which we omit for brevity.)

Prove the converse of what was shown in class. That is, prove that if $(\mathbf{w}^*, \boldsymbol{\alpha}^*)$ satisfies Eq. (4), then Eqs. (1), (2) and (3) are also satisfied. You should not assume anything special about L (such as convexity), but you can assume all of the relevant minima and maxima exist.

b. [10] Let $a_1, \ldots, a_n$ be nonnegative real numbers, not all equal to zero, and let $b_1, \ldots, b_n$ and c all be positive real numbers. Use the method of Lagrange multipliers to find the values of $x_1, \ldots, x_n$ which minimize

$$-\sum_{i=1}^{n} a_i \ln x_i$$

subject to the constraint that

$$\sum_{i=1}^{n} b_i x_i \leq c.$$

Show how this implies that relative entropy is nonnegative.

Problem 2

Consider the problem of learning with expert advice when one of the experts gives perfect predictions. On some round t, let q be the fraction of surviving experts that predict 1. (A surviving expert is one that has not made any mistakes so far.) In class, we talked about the halving algorithm which predicts with the majority vote of the expert predictions, and we talked about the randomized weighted majority algorithm (with β set to zero) which predicts with one randomly selected expert.

In general, we can predict 1 with probability $F(q)$ and 0 with probability $1 - F(q)$ for some function F. For instance, for the halving algorithm, $F(q)$ is 1 if $q > 1/2$ and 0 if $q < 1/2$ (and arbitrary if $q = 1/2$). For the randomized weighted majority algorithm (again, with $\beta = 0$), $F(q) = q$.

Consider now a function $F : [0,1] \to [0,1]$ satisfying the following property:

$$1 + \frac{\lg q}{2} \le F(q) \le -\frac{\lg(1-q)}{2}. \tag{5}$$

a. [15] Suppose we run an on-line learning algorithm that uses a function F satisfying (5) as described above. Show that the expected number of mistakes made by the learning algorithm is at most $(\lg N)/2$, where N is the number of experts.

b. [10] Show that the function

$$F(q) = \frac{\lg(1-q)}{\lg q + \lg(1-q)}$$

has range $[0,1]$ and satisfies (5). (At the endpoints, we define $F(0) = 0$ and $F(1) = 1$ to make F continuous, but you *don't* need to worry about these.)

c. [10] **(Optional – for extra credit)** Suppose now that there are $k \ge 2$ possible outcomes rather than just 2. In other words, the outcome y_t is now in the set $\{1, \dots, k\}$ (rather than $\{0,1\}$ as we have considered up until now), and likewise, both experts and the learning algorithm make predictions in this set. Assume one of the experts makes perfect predictions. On some round t, let q_j be the fraction of surviving experts predicting outcome $j \in \{1, \dots, k\}$. Suppose that the learning algorithm predicts each outcome j with probability

$$\frac{\lg(1-q_j)}{\sum_{i=1}^{k} \lg(1-q_i)}.$$

Show that the expected number of mistakes of this learning algorithm is at most $(\lg N)/2$.

Problem 3

[15] For this problem, let us suppose that labels, outcomes, expert/hypothesis predictions, etc. are all defined over the set $\{-1,+1\}$ rather than $\{0,1\}$. Since this does not change what it means for the learner or an expert to make a mistake, this has no effect on any of the results we have discussed regarding online mistake bounds.

Let $\mathcal{H}$ be a finite space of hypotheses $h : \mathcal{X} \to \{-1,+1\}$, and let $S = \langle x_1, \dots, x_m \rangle$ be any sequence of m distinct points in $\mathcal{X}$. Prove that the empirical Rademacher complexity of $\mathcal{H}$ satisfies

$$\hat{\mathcal{R}}_S(\mathcal{H}) \le O\left(\sqrt{\frac{\ln|\mathcal{H}|}{m}}\right)$$

by applying our analysis of online algorithms for learning with expert advice to an appropriately constructed sequence of expert prediction ξ_i and outcomes y. Give a bound with explicit constants.

Note that this bound was earlier stated without proof in class (see Theorem 2 in the scribe notes for lecture #10), and is also a special case of Theorem 3.3 in the book, although with possibly weaker constants. **Extra credit** [5] will be given for obtaining a bound of $\sqrt{(2\ln|\mathcal{H}|)/m}$, that is, with the constant that was stated in class.