Computational and Statistical Learning theory
Assignment 4

Due: March 16th
Email solutions to : karthik *at* ttic *dot* edu

Definitions/Notation

Throughout we assume that the set $\mathcal{W}$ is a closed convex subset of a Banach space $\mathcal{B}$ equipped with norm $\|\cdot\|$. Let $\|\cdot\|_*$ be the dual norm.

Definition 1. *A function $F : \mathcal{W} \mapsto \mathbb{R}$ is said to be σ-strongly convex w.r.t. norm $\|\cdot\|$ on $\mathcal{W}$ if for any $w, w' \in \mathcal{W}$,*

$$F(w) \geq F(w') + \langle \nabla F(w'), w - w' \rangle - \frac{\sigma}{2}\|w - w'\|^2$$

Definition 2. *Given a strictly convex function $F : \mathcal{W} \mapsto \mathbb{R}$, the Bregman Divergence of the function is given by*

$$\Delta_F(w, w') := F(w) - F(w') - \langle \nabla F(w), w - w' \rangle$$

Definition 3. *Given a convex function $F : \mathcal{W} \mapsto \mathbb{R}$, its dual F^* is defined as*

$$F^*(x) = \sup_{w \in \mathcal{W}} \{\langle x, w \rangle - F(w)\}$$

You might find the following property of Bregman divergences useful :

$$\nabla F^* = (\nabla F)^{-1}$$

For any convex function ℓ on $\mathcal{W}$ we shall use the notation $\partial \ell(w)$ to represent the set of sub-gradients of ℓ at point w or in other words the set

$$\partial \ell(w) = \{\lambda : \forall w' \in \mathcal{W},\ \ell(w') - \ell(w) \geq \langle \lambda, w' - w \rangle\}$$

Any convex function ℓ has at least one sub-gradeint at all points and if the function is differentiable at a point say w there is exactly one sub-gradient at that point given by gradient $\nabla \ell(w)$.

Problems

1. **Lower Bound for Perceptron :**
   For any $\gamma > 0$, let $d \geq \frac{1}{\gamma^2}$ and $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ and $\mathcal{Y} = \{\pm 1\}$. Show that for any online learning algorithm, there exists a sequence of instances $(x_1, y_1), \ldots, (x_m, y_m)$ which is separable by a margin of γ by some linear separator with ℓ_2 norm bounded by 1, such that the online algorithm makes at least $\lfloor \frac{1}{\gamma^2} \rfloor$ mistakes on this sample. This shows that the perceptron bound is tight.

   Hint : Pick appropriate m (depending on γ) and provide instances adversarially so that the algorithm makes a mistake on every round. However show that the selected instances are separable by a linear separator of norm 1 with a margin of at least γ.

2. **Mirror Descent Guarantee (optional):**
   Let $\mathcal{W}$ be a convex closed subset of some Banach space equipped with norm $\|\cdot\|$. Let $F : \mathcal{W} \mapsto \mathbb{R}^+$ be some non-negative σ-strongly convex function on $\mathcal{W}$. Further assume that $B^2 = \sup_{w \in \mathcal{W}} F(w)$. Consider the mirror descent update given by

   $$w_{t+1} \leftarrow \underset{w \in \mathcal{W}}{\operatorname{argmin}} \langle \eta \lambda_t - \nabla F(w_t), w \rangle + F(w)$$

   where $\lambda_t \in \partial \ell_t(w_t)$. The above update can equivalently be given by the two step update

   $$w'_{t+1} \leftarrow \nabla F^* \left(\nabla F(w_t) - \eta \lambda_t \right) \quad \text{and} \quad w_{t+1} \leftarrow \underset{w \in \mathcal{W}}{\operatorname{argmin}} \Delta_F(w, w'_{t+1})$$

   We would like to prove that by selecting η appropriately, for any sequence $\ell_1, \ldots, \ell_m$ of L-Lipschitz convex functions chosen by the adversary, the regret of the mirror descent algorithm using F is bounded as :

   $$\frac{1}{m} \sum_{t=1}^{m} \ell_t(w_t) - \inf_{w \in \mathcal{W}} \frac{1}{m} \sum_{t=1}^{T} \ell_t(w) \leq \sqrt{\frac{8 L^2 B^2}{\sigma \, m}}$$

   We shall prove the above statement by taking the following steps :

   (a) For any $w, w', w'' \in \mathcal{W}$ prove that

   $$\langle \nabla F(w') - \nabla F(w''), w'' - w \rangle = \Delta_F(w, w') - \Delta_F(w, w'') - \Delta_F(w'', w')$$

   (b) Prove that for any $\eta > 0$, any $w \in \mathcal{W}$ and any $t \in [m]$,

   $$\eta \langle \lambda_t, w_t - w \rangle \leq \Delta_F(w, w_t) - \Delta_F(w, w_{t+1}) + \frac{\eta^2}{2\sigma} \|\lambda_t\|_*^2$$

   Hint: You will have to use mirror descent update rule, the equation you proved in step 1 and the inequality that $\langle w, \lambda \rangle \leq \|w\| \|\lambda\|_*$ along with the fact that for any two numbers $a, b \in \mathbb{R}$, $ab \leq \frac{a^2}{2} + \frac{b^2}{2}$. Also notice that the definitions of strong convexity of F and Bregman divergence directly imply that for any $w, w' \in \mathcal{W}$,

   $$\Delta_F(w', w) \geq \frac{\sigma}{2} \|w - w'\|^2$$

(c) Use this (along with convexity and Lipschitz property of ℓ_t's) to prove that for appropriately chosen η,

$$\frac{1}{m}\sum_{t=1}^{m}\ell_t(w_t) - \inf_{w\in\mathcal{W}}\frac{1}{m}\sum_{t=1}^{T}\ell_t(w) \le \sqrt{\frac{8L^2B^2}{\sigma\, m}}$$

Hint : There is a telescoping sum involved.

Challenge Problems

1. **Regularized ERM :**
For the same set up as the Mirror Descent question in problem 2, that is $\mathcal{W}$ is a convex closed subset of some Banach space. $F : \mathcal{W} \mapsto \mathbb{R}^+$ is some non-negative σ-strongly convex function on $\mathcal{W}$ and $B^2 = \sup_{w\in\mathcal{W}} F(w)$. Prove that for appropriate choice of regularization parameter β, the learning rule given by F-regularized ERM :

$$\tilde{w} = \operatorname*{argmin}_{w\in\mathcal{W}} \frac{1}{m}\sum_{i=1}^{m}\ell_i(w) + \beta F(w)$$

enjoys the following bound on risk for the statistical convex optimization problem :

$$R(\tilde{w}) \le \inf_{w\in\mathcal{W}} R(w) + \sqrt{\frac{8L^2B^2}{\sigma m}}\ .$$

2. ℓ_1 **Regularized Learning Lower Bound :**
For any $\gamma > 0$, for appropriately chosen d and $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_\infty \le 1\}$ and $\mathcal{Y} = \{\pm 1\}$. Show that for any online learning algorithm, there exists a sequence of instances $(x_1, y_1), \ldots, (x_m, y_m)$ which is separable by a margin of γ by some linear separator with ℓ_1 norm bounded by 1, such that the number of mistakes made by the online algorithm say M is lower bounded as

$$M \ge \Omega\left(\max\left\{\log d, \frac{1}{\gamma^2}\right\}\right)$$

Research Problems

1. ℓ_1 **Regularization Lower Bound :**
Given $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_\infty \le 1\}$ and $\mathcal{Y} = \{\pm 1\}$, the best upper bound on number of mistakes M we know (using Winnow algorithm) when there exists a linear separator with ℓ_1 norm bounded by 1 which separates the examples with margin γ is

$$M \le O\left(\frac{\log d}{\gamma^2}\right)$$

Challenge problem 2 only gives lower bound that is best of $\frac{1}{\gamma^2}$ and $\log d$. Can you show a lower bound of form

$$M \geq \Omega\left(\frac{\log d}{\gamma^2}\right)$$

or show the tightest possible lower and upper bounds?

2. **Regularization and Statistical Learning :**
Consider any stochastic convex optimization problem where objective $r : \mathcal{W} \times \mathbb{Z} \mapsto \mathbb{R}$ is convex and L-lipschitz in its first argument for all $z \in \mathbb{Z}$. The learner is provided with sample $S = \{z_1, \ldots z_m\}$ drawn iid from some unknown distribution $\mathcal{D}$ and is expected to pick some $\tilde{w} \in \mathcal{W}$.based on this sample. Recall that the problem is defined to be learnable if the learner can pick a learning algorithm that returns $\tilde{w}$ that satisfies,

$$\mathbb{E}_S\left[R(\tilde{w}) - \inf_{w \in \mathcal{W}} R(w)\right] \rightarrow 0$$

Prove or disprove the following statement :

The problem is learnable if and only if there exists a regularizer function $F : \mathcal{W} \mapsto \mathbb{R}$ such that F-regularized ERM rule given by

$$\tilde{w} = \underset{w \in \mathcal{W}}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} r(w, z_i) + \beta F(w)$$

with appropriate β (depending on m) provides for a successful learning rule.

The motivation for this question is that all the cases of stochastic convex optimization problems we know that are statistically learnable are learnable because of uniform convergence (in which case $\beta = 0$ and F can be arbitrary) or when the problem is learnable online in which case we can argue that there always exists a regularizer that has nice properties which can be used for learning. Is there any other type of problem?