Computational and Statistical Learning Theory

Problem set 7

Due: November 21th

Please send your solutions to `learning-submissions@ttic.edu`

Notations/Definitions

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

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

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. **Direct analysis of non-separable perceptron :**

   (a) Recall the perceptron rule : if $y\langle w, x \rangle \leq 0$, then add yx to w.

Instead of assuming the existence of w s.t. for all t, $y_t\langle w, x_t\rangle > 1$ (setting $\gamma = 1$), we will derive a mistake bound that bounds the number of mistakes the (standard) perceptron makes in terms of best possible total hinge loss on the sequence.

For any sequence $(x_t, y_t)_{t=1\ldots m}$, where $\|x_t\| \leq 1$ and $y_t \in \{\pm 1\}$, let M_m be the number of mistakes made by the perceptron:

$$M_m = \{t = 1 \ldots m | y_t\langle w_t, x_t\rangle \leq 0\}$$

For any w^*, let H^*_m be the total hinge loss of w^* on the sequence:

$$H^*_m = \sum_{t=1}^{m} [1 - y_t\langle w^*, x_t\rangle]_+$$

Prove the following:

$$M_m \leq H^*_m + \|w^*\|^2 + \|w^*\|\sqrt{H^*_m}$$

Hint: follow the perceptron analysis as in class: Bound $\|w_{t+1}\|^2$ from above in terms of M_t. Then bound $\langle w^*, w_{t+1}\rangle$ from below in terms of both M_t and H^*_t. Combine the two bounds and solve a quadratic equation to obtain the bound on M_m.

(b) Use an online-to-batch conversion to obtain a learning rule A for which, with high probability, for every w^* with $\|w^*\|_2 \leq B$:

$$L_{01}(A(S)) \leq L^*(w) + O(B^2/m + \sqrt{B^2 L^*/m})$$

where $L^* = L_{\text{hinge}}(w^*)$. State the learning rule explicitly and prove the learning guarantee. Is this a proper learning rule? What form of predictors does it output?

3. Prove that $\Psi_p(w) = \frac{1}{2}\|w\|_p^2$ is $(p-1)$-strongly convex w.r.t. the norm $\|w\|_p$.

4. Consider the entropy regularizer $\Psi(w) = \log(d) - \sum_i w[i]\log(w[i])$

   (a) Prove that as long as $w' \geq 0$, then $\operatorname{argmin}_{w \in W} d_{\Psi(w,w')} = w'/\|w\|_1$ where $W = \{w \in R^d | w[i] \geq 0, \|w\|_1 = 1\}$ is the simplex

   (b) Calculate the parameter of strong convexity of $\Psi(w)$ over $\{w \in R^d | \|w\|_1 = B, w[i] \geq 0\}$.

   (c) Suggest a non-negative regularizer (based on the entropy regularizer) which is 1-strongly convex over $W = \{w \in R^d | \|w\|_1 \leq B\}$. Calculate $\sup_{w \in W} \Psi(w)$, derived in closed form the mirror descent update, and state the corresponding learning guarantee.

Challenge Problems

1. ℓ_1 **Regularized Learning Lower Bound :**
   For any $\gamma > 0$, for appropriately chosen d and $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_\infty \leq 1\}$ 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_1 norm bounded by 1, such that the number of mistakes made by the online algorithm say M is lower bounded as

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

2. Instead of the mirror descent update, consider updates of the form:

$$w_{t+1} = \operatorname*{argmin}_{w \in W} l(w, z_t) + \frac{1}{\eta} d_\Psi(w, w_t)$$

prove a regret guarantee for the above update that is similar to the mirror descent guarantee

3. Consider the follow-the-regularized-leader rule:

$$w_{t+1} = \arg min_{w \in W} \frac{1}{t} \sum_{i=1\ldots t} l(w, z_i) + \lambda \Psi(w)$$

(a) With $\lambda = 0$, we obtain the "follow the leader" learning rule. Prove that for $l(w, (x, y)) =$ hinge$(y\langle w, x\rangle)$, $W = \{w | \|w\|_2 \leq B\}$, $X = \{x | \|x\|_2 < 1\}$, we may have non-diminishing regret using this rule. Ie that there is some c s.t. for all m, there is a sequence of length m with average regret $\geq c$. Note that if there are multiple minimizers for the argmin, it is enough to show that there exists a choice of minimizers that yields non-diminishing regret.

(b) Prove a regret guarantee similar to the MD guarantee for follow-the-regularized leader with an appropriate choice of lambda. What is an appropriate choice?

Research Problems

1. ℓ_1 **Regularization Lower Bound :**
Given $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_\infty \leq 1\}$ 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

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

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,

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

Prove or disprove the following statement :

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

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

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 Ψ 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?