Computational and Statistical Learning Theory

Problem set 4

Due: Friday, December 2nd

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

Notation:

- Input space: $\mathcal{X}$
- Label space: $\mathcal{Y} = \{\pm 1\}$
- Sample: $(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X}$
- Hypothesis Class: $\mathcal{H}$
- Risk: $L_{\mathcal{D}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{1}_{h(x)\neq y}\right]$
- Empirical Risk: $L_S(h) = \frac{1}{m}\sum_{(x,y)\in S}\mathbf{1}_{h(x)\neq y}$

1. **Perceptron**:

   (a) 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.

   (b) Direct analysis of non-separable perceptron:

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

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

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 $|M_m|$.

ii. Use an online-to-batch conversion to obtain a learning rule A that output a linear predictor, for which, with high probability, for every w^* with $\|w^*\|_2 \leq B$:

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

where $L^* = L_{\text{hinge}}(w^*)$. State the learning rule explicitly and prove the learning guarantee.

2. **Stability:**
Prove replace-one stability of RERM with $\Psi(w)$ regularization, where $\Psi(w)$ is α-strongly convex w.r.t. $\|w\|$ and $\ell(w, z)$ is G-Lipschitz w.r.t $\|w\|$. That is, there is a quantity β, such that for all $z_1, ..., z_m$ and z'_i:

$$|\ell(A(z_1, ..., z_m), z_i) - \ell(A(z_1, ..., z'_i, ..., z_m), z_i)| \leq \beta,$$

specify β in terms of m, α, G.

3. Using $\Psi(w) = \|w\|_p^2$ as a regularizer,

(a) Prove $\Psi(w)$ is α-strongly convex w.r.t. $\|w\|_p$.

(b) Calculate

i. Link function $\nabla\Psi(w)$.

ii. Link function $\nabla\Psi^{-1}(\nu)$

iii. Bregman divergence $D_\Psi(w\|w')$

(c) Consider supervised learning, with linear predictor, hinge loss, $\|\phi(x)\|_\infty \leq 1$, $\phi(x) \in \mathbb{R}^d$, and learning the hypothesis class $\mathcal{H} = \left\{w \;\middle|\; \|w\|_1 \leq B\right\}$. We would like to learn this hypothesis class using L-FTRL with the regularizer $\Psi(w) = \|w\|_p^2$.

i. Show that for an appropriate value of p, we get online regret $O(\sqrt{B^2 \log(d)/m})$. State the value of p exactly, specify the learning rule, and in particular what sequence of regularization parameters you would use, and give the resulting regret bound exactly (not using big-O notation).

ii. Derive explicit pseudo-code for the update

iii. Instead consider using non-linearized online mirror descent. Derive explicit pseudo-code for the update

4. Consider "following the leader" learning rule as:

$$w_{t+1} = \arg\min_{w \in W} \frac{1}{t} \sum_{i=1}^{t} \ell(w, (x_i, y_i)).$$

where $W = \{w | \|w\|_2 \leq B\}$ and $\max_i \|x_i\| \leq 1$.

(a) Prove if $\ell(\cdot, \cdot)$ is hinge loss, we may have non-diminishing regret using this rule, i.e. that there is some c s.t. for all t, there is a sequence of length t 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 if $\ell(\cdot, \cdot)$ is squared loss, i.e. $\ell(w, (x, y)) = \frac{1}{2}(y - w^T x)^2$, we may have non-diminishing regret using this rule.

(c) Prove if $\ell(\cdot, \cdot)$ is squared loss, and x are drawn from multivariate normal distribution with zero mean and identity covariance, then we have sub-linear regret.