E1 245 - Online Prediction and Learning, Aug-Dec 2018
Homework #3

1. *Strong convexity of entropy (Pinsker's inequality)*

   (a) Prove that the negative entropy function $R(x) = \sum_i x_i \log x_i$ over the 2 dimensional simplex $\Delta_2 := \{(x_1, x_2) : x_i \geq 0, x_1 + x_2 = 1\}$ is 1-strongly convex with respect to the $||\cdot||_1$ norm.

   (b) Prove the same statement when $d = 2$ is replaced with a general positive integer $d \geq 2$.
   Hint: One way is to find a reduction to the $d = 2$ case. Let x and y be two vectors in Δ_d. Let $A := \{i : x_i \geq y_i\}$ be the coordinates where x dominates y. Can you find two new vectors x_A and y_A in Δ_2 so that $||x - y||_1 = ||x_A - y_A||_1$ and carry on?

2. *Exponential Weights as FTRL and OMD*

   (a) Show that Follow The Regularized Leader (FTRL) on the simplex $\Delta_N := \left\{(x_1, \ldots, x_N) : \sum_{i=1}^N x_i = 1, \forall i\ x_i \geq 0\right\}$ with the entropic regularizer[1] $R_\eta(x) := \frac{1}{\eta}\sum_{i=1}^N x_i \log x_i$, and linear loss functions $f_t(x) = \langle z_t, x\rangle$, is equivalent to running the Exponential Weights algorithm on N experts with loss vectors $\{z_t\}_{t\geq 1}$ and parameter η.
   Hint: You can derive this directly from first principles and the definition of the FTRL rule. An alternative way is by using (a) the equivalence between FTRL and (unconstrained minimization + Bregman projection) proven in class, and (b) observing that Bregman projection wrt the regularizer R onto Δ_N is equivalent to scaling by the $||\cdot||_1$ norm.

   (b) Show that Active Online Mirror Descent on the simplex $\Delta_N := \left\{(x_1, \ldots, x_N) : \sum_{i=1}^N x_i = 1, \forall i\ x_i \geq 0\right\}$ with the entropic regularizer $R_\eta(x) := \frac{1}{\eta}\sum_{i=1}^N x_i \log x_i$, and linear loss functions $f_t(x) = \langle z_t, x\rangle$, is equivalent to running the Exponential Weights algorithm on N experts with loss vectors $\{z_t\}_{t\geq 1}$ and parameter η.

3. *Analysing Exponential Weights as OMD*

   (a) Prove the following result. Suppose (active) OMD is run on the convex decision set $\mathscr{K}$ with a Legendre function R, where R is α-strongly convex with respect to some norm $||\cdot||$ on $\mathscr{K}$, $R(x) - R(w_1) \leq B^2\ \forall x \in \mathscr{K}$, and the gradients of the loss functions are at most G in the dual[2] norm $||\cdot||_*$. Then, with a step size $\eta := \frac{G}{B}\sqrt{\frac{2}{T}}$, the T-round regret of OMD is at most $BG\sqrt{\frac{2T}{\alpha}}$.
   Hint: In the regret bound for active OMD in class, upper bound the term $D_R(w_t, w'_{t+1}) - D_R(w_{t+1}, w'_{t+1})$.

   (b) Using this and the previous exercises, argue an appropriate regret bound for the Exponential weights algorithm run on the simplex Δ_d, and with linear loss functions having coefficients in $[0, 1]$.

4. *Fenchel duality and Bregman divergence*
   For each of the following functions defined on $\mathbb{R}^d$, compute its gradient, Fenchel dual, gradient of the Fenchel dual, and the Bregman divergences of itself and its Fenchel dual.

   (a) $F_1(x) = e^{x_1} + \cdots + e^{x_d}$.

   (b) $F_2(x) = \log(e^{x_1} + \cdots + e^{x_d})$.

   (c) $F_3(x) = \frac{1}{2}||x||_p^2$, $p \in [1, \infty]$.

5. *Bregman projection*
   Show that the projection of $y \in (0, \infty)^d$ onto the probability simplex Δ_d, with respect to the Bregman divergence induced by the generalized negative entropy ($R(x) := \sum_{i=1}^2 x_i \log x_i - x_i$), is simply its normalization, i.e., $\Pi_{\Delta_d}^{D_R}(y) = y/||y||_1$.

[1] $0 \log 0$ is defined to be 0.
[2] Recall that for a norm $||\cdot||$ in $\mathbb{R}^d$, its dual norm is defined by $||y||_* := \max_{x:||x||=1} x^T y$.