Homework 3 (Due Date Sunday April 19 11.59pm)

Please write the following on your homework:

- Name
- Collaborators (write none if no collaborators)
- Source, if you obtained the solution through research, e.g. through the web.

While you may collaborate, you *must write up the solution yourself.* While it is okay for the solution ideas to come from discussion, it is considered as plagiarism if the solution write-up is highly similar to your collaborator's write-up or to other sources. You solution should be submitted to IVLE workbin. Scanned handwritten solutions are acceptable but must be legible.
Late Policy: A late penalty of 20% per day will be imposed (no submission accepted after 5 late days) unless prior permission is obtained.

---

1. **Perceptron Algorithm as Online Convex Optimization**
   The perceptron algorithm does online learning of a linear threshold function, i.e. $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{-1, 1\}$ and the prediction $\hat{y}_t = \text{sign}(\langle \mathbf{w}^{(t)}, \mathbf{x}_t \rangle)$. Each time the algorithm makes a mistake, it updates the weight vector $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \eta y_t \mathbf{x}_t$. If the algorithm makes the correct prediction, the weight vector is unchanged.

   (a) Argue that the perceptron algorithm is doing online gradient descent with the loss function $\ell_t(\mathbf{w}^{(t)}, \mathbf{x}_t, y_t) = 0$ for rounds when the prediction is correct, and $\ell_t(\mathbf{w}^{(t)}, \mathbf{x}_t, y_t) = \max\{0, 1 - y_t \langle \mathbf{w}^{(t)}, \mathbf{x}_t \rangle\}$ when the prediction is incorrect.

   (b) Argue that the total loss upper bounds the total number of mistakes.

   (c) Assume that there exists $\mathbf{w}^*$ such that $y_t \langle \mathbf{w}^*, \mathbf{x}_t \rangle \geq 1$ for all t (i.e. correct with margin at least 1), and argue that $\mathbf{w}^*$ has zero total loss.

   (d) Using SSBD Lemma 14.1, argue that the number of mistakes $M \leq R^2 \|\mathbf{w}^*\|^2$, where $R = \max \|\mathbf{x}_t\|$.

   (e) Further argue that the mistake bound still holds when η is set to 1.

2. **Switching Predictors**
   Assume that we have a finite class H of expert binary predictors and we have to do T rounds of online predictions. Instead of assuming that there is a perfect predictor in H, we assume that we have a sequence of predictors $h_1, \ldots, h_{k+1}$ from H where we switch from h_i to h_{i+1} at time t_i such that the switching sequence of predictors give perfect predictions. We call such a predictor a k-switching predictor. We would like to apply a version of the Halving algorithm to the problem where a k-switching predictor which makes perfect prediction exists (a version of the weighted majority algorithm can also be used when the k-switching predictor can make mistakes).

(a) We first modify the Halving algorithm to handle weighted hypotheses. Assume that hypothesis h is given a weight w_h such that $\sum_{h \in H} w_h = 1$. In the modified Halving algorithm, we remove a hypothesis from H whenever it makes a mistake (set its weight to 0), and predict using $\arg\max_{r \in \{0,1\}} \sum_{h \in H, h(x)=r} w_h$, i.e. we predict using the label which agrees with the weighted majority of surviving hypotheses. Show that the mistake bound for this algorithm is no more than $\log_2 1/w_{h^*}$ where h^* is a predictor that does not make any error.

(b) We now assume that there exists a k-switching predictor that makes no mistake. Assume that we use the following weighting scheme for each k switching predictor (predictor with k switches): $\frac{1}{|H|(|H|-1)^k} p^k (1-p)^{T-k-1}$, where p is a user-selected parameter. We will argue that the sum of the weights of all such switching predictors (with $k = 1, \ldots, T-1$) is 1 by mathematical induction.

   i. Argue that the statement is true for $T = 1$.

   ii. Assume that the statement is true for predictor sequences of length $T-1$, then show that the statement is true for predictor sequences of length T.

(c) Argue that the number of errors made by the modified Halving algorithm is not more than $\log_2 |H| + k \log_2(|H|-1) + k \log_2 1/p + (T-k-1) \log_2 1/(1-p)$ (approximately $k(\log_2 |H| + \log_2 1/p)$ when p is small) when there is a k-switching predictor that predicts the sequence prefectly.

(d) Give an efficient algorithm for running the modified Halving algorithm on all possible k-switching predictors. The algorithm should run in time $O(|H|^2)$ (or $O(|H|)$ after some optimization) per iteration. (Hint: Try constructing an algorithm by modifying the inductive proof in part (b).)