CSCI 678: Theoretical Machine Learning
Quiz 3 (15 minutes)

Fall 2026, Instructor: Haipeng Luo

NAME: _____

5 multiple-choice questions (10 points). There is ONLY ONE correct answer for each question.

1. Consider learning over a finite class of binary classifiers $\mathcal{F} \subset \{-1,+1\}^{\mathcal{X}}$ with respect to the 0-1 loss and assume that the realizable assumption holds, that is, $\inf_{f\in\mathcal{F}} \sum_{t=1}^{n} \mathbf{1}\{f(x_t) \neq y_t\} = 0$. Which of the following statements about Halving and Hedge with learning rate $\eta = 1/2$ is correct? C

   (A) Both are improper algorithms that always make at most $\mathcal{O}(\ln|\mathcal{F}|)$ mistakes.

   (B) Both are proper algorithms that always make at most $\mathcal{O}(\ln|\mathcal{F}|)$ mistakes.

   (C) Halving is an improper algorithm that always makes at most $\mathcal{O}(\ln|\mathcal{F}|)$ mistakes.

   (D) Hedge is a proper algorithm that always makes at most $\mathcal{O}(\ln|\mathcal{F}|)$ mistakes.

2. Which of the following statements about the Perceptron algorithm is correct? D

   (A) Perceptron always makes at most a constant number of mistakes.

   (B) Perceptron updates $\theta \leftarrow \theta + y_t x_t$ every round.

   (C) Perceptron updates $\theta \leftarrow \theta + y_t x_t$ only when it makes a correct prediction in that round.

   (D) Perceptron updates $\theta \leftarrow \theta + y_t x_t$ only when it makes an incorrect prediction in that round.

3. Suppose that $\mathcal{X} = B_2^d$ (d-dimensional unit ℓ_2-norm ball) and $\boldsymbol{x}$ is an $\mathcal{X}$-valued tree of depth n that is α-shattered by $\mathcal{F} = \left\{f_\theta(x) = \langle\theta, x\rangle \mid \theta \in B_2^d\right\}$ with witness $\boldsymbol{y}$ (a $[-1,+1]$-valued tree). Now consider a binary classification dataset $(z_1,\epsilon_1),\ldots,(z_n,\epsilon_n)$ where $z_t = (\boldsymbol{x}_t(\epsilon_{1:t-1}), \boldsymbol{y}_t(\epsilon_{1:t-1})) \in \mathbb{R}^{d+1}$ and $\epsilon_t \in \{-1,+1\}$ is arbitrary. What is the largest value of $\gamma > 0$ such that there must exist $\theta^\star \in B_2^{d+1}$ satisfying $\epsilon_t \langle\theta^\star, z_t\rangle \geq \gamma$ for all $t = 1,\ldots,n$? B

   (A) $\frac{\alpha}{4}$

   (B) $\frac{\alpha}{2\sqrt{2}}$

   (C) $\frac{\alpha}{2}$

   (D) $\frac{\alpha}{\sqrt{2}}$

Algorithm 1: Winnow

Let $\theta = \mathbf{0}$. For $t = 1, \ldots, n$:

- Receive x_t and predict $s_t = \text{sign}(\langle x_t, g(\theta) \rangle)$ where $g(\theta) = \exp(\eta\theta) - \exp(-\eta\theta)$.
- Receive $y_t \in \{-1, +1\}$. If $y_t \neq s_t$, update $\theta \leftarrow \theta + y_t x_t$.

Algorithm 2: Simplified Winnow

Let $\theta = \mathbf{0}$. For $t = 1, \ldots, n$:

- Receive x'_t and predict $s_t = \text{sign}(\langle x'_t, g(\theta) \rangle)$ where $g(\theta) = \exp(\eta\theta)$.
- Receive $y_t \in \{-1, +1\}$. If $y_t \neq s_t$, update $\theta \leftarrow \theta + y_t x'_t$.

4. In HW3, you have seen the Winnow algorithm (Algorithm 1) and its simplified version (Algorithm 2). Which of the following configurations makes the two algorithms produce the exact same prediction s_t for all t? A

   (A) $x'_t = (x_t, -x_t)$

   (B) $x'_t = (-x_t, -x_t)$

   (C) $x'_t = (-x_t, x_t)$

   (D) $x'_t = -x_t$

5. Continuing from the last question and using the form of x'_t that makes the two algorithms equivalent, suppose that $\theta^\star$ is such that $y_t \langle \theta^\star, x_t \rangle \geq \gamma$ for all t. Which of the following formulas for θ satisfies $y_t \langle \theta, x'_t \rangle \geq \gamma$ for all t? (Below, $|\theta^\star|$ represents taking coordinate-wise absolute values of $\theta^\star$.) D

   (A) $\theta = |\theta^\star|$

   (B) $\theta = (\theta^\star, -|\theta^\star|)$

   (C) $\theta = (|\theta^\star|, -|\theta^\star|)$

   (D) $\theta = \frac{1}{2}(|\theta^\star| + \theta^\star, |\theta^\star| - \theta^\star)$