M2 MAF - MATHEMATICS OF MACHINE LEARNING
FINAL EXAM - MARCH 3RD, 2016

Problem 1

Preliminary : a Bernstein Bound.
For every $u \in \mathbb{R}$, let $\phi(u) = e^u - u - 1$, and let

$$\psi(u) = \begin{cases} \frac{\phi(u)}{u^2} & \text{if } u \neq 0 \ , \\ \frac{1}{2} & \text{if } u = 0 \ . \end{cases}$$

(1) Show that ψ is an increasing function on $\mathbb{R}$.

(2) Let X be random variable almost-surely upper-bounded by 1. Show that

$$\forall \lambda > 0, \qquad \mathbb{E}\left[e^{\lambda X}\right] \leqslant 1 + \lambda \mathbb{E}[X] + \mathbb{E}[X^2]\phi(\lambda) \ ,$$

and hence that

$$\ln \mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right] \leqslant \mathbb{E}[X^2]\, \phi(\lambda) \ .$$

(3) Show that if Y is a random variable almost-surely upper-bounded by $b > 0$,

$$\forall \lambda > 0, \qquad \ln \mathbb{E}\left[e^{\lambda(Y - \mathbb{E}[Y])}\right] \leqslant \frac{\mathbb{E}[Y^2]}{b^2}\, \phi(\lambda b) \ .$$

(4) Show that if Z is a random variable almost-surely upper-bounded by 0,

$$\forall \lambda > 0, \qquad \ln \mathbb{E}\left[e^{\lambda(Z - \mathbb{E}[Z])}\right] \leqslant \frac{\mathbb{E}[Z^2]\lambda^2}{2} \ .$$

Regret Bound for Individual Sequences.

In this section, we consider the following online learning problem, which can be seen as a particular case of prediction of individual sequences with the linear loss :

At each round $t \in \mathbb{N}^*$,

— The statistician chooses a weight vector $p_t \in \Delta_K = \left\{u \in \mathbb{R}_+^K : \sum_{i=1}^K u_i = 1\right\}$ as a function of the past data $\ell_1, \ldots, \ell_{t-1}$.

— The statistician observes the vector $\ell_t = (\ell_{i,t})_{1 \leqslant i \leqslant K} \in [0,1]^K$ and incurs the loss $p_t \cdot \ell_t = \sum_{i=1}^K p_{i,t}\ell_{i,t}$.

We consider the EWA algorithm, whose weight vectors $p_t = \left(p_{i,t}\right)_{1 \leqslant i \leqslant K}$, are defined as follows ($\eta > 0$ is a parameter of the algorithm) :

$$p_{i,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell_{i,s}\right)}{\sum_{j=1}^K \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{j,s}\right)} \ , \quad 1 \leqslant i \leqslant K \ .$$

The goal of this section is to prove that the EWA algorithm satisfies the regret bound (I) below.

(5) We set $L_{i,0} = 0$ and $L_{i,t} = \sum_{s=1}^{t} \ell_{i,s}$ for all $t \geqslant 1$ and $i \in \{1, \dots, K\}$. We also define $W_t = \frac{1}{K} \sum_{i=1}^{K} e^{-\eta L_{i,t-1}}$ for all $t \geqslant 1$. Prove that

$$\ln \frac{W_{T+1}}{W_1} \geqslant -\eta \min_{1 \leqslant i \leqslant K} L_{i,T} - \ln K \ .$$

(6) Show that

$$\ln \frac{W_{T+1}}{W_1} \leqslant -\eta \sum_{t=1}^{T} p_t \cdot \ell_t + \frac{\eta^2}{2} \sum_{t=1}^{T} \sum_{i=1}^{K} p_{i,t} \, \ell_{i,t}^2 \ .$$

(7) Prove that the EWA algorithm satisfies the following regret bound : for all $T \geqslant 1$ and all $\ell_1, \dots, \ell_T \in [0,1]^K$,

$$\text{(I)} \qquad \sum_{t=1}^{T} p_t \cdot \ell_t \leqslant \min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \ell_{i,t} + \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{K} p_{i,t} \, \ell_{i,t}^2 \ .$$

Regret Bound with Bandit Feedback.

In this section, we consider another online learning problem known as *adversarial multi-armed bandit setting*, which unfolds as follows :

> Initialization : a sequence of deterministic loss vectors $\ell_1, \ell_2, \dots \in [0,1]^K$ is chosen by an adversary but not revealed to the statistican (yet).
>
> At each round $t \in \mathbb{N}^*$,
>
> — The statistician chooses a weight vector $p_t \in \Delta_K$ as a function of the past data $\ell_{I_1,1}, \dots, \ell_{I_{t-1},t-1}$, and draws $I_t \in \{1, \dots, K\}$ at random such that
>
> $$\forall i \in \{1, \dots, K\} \ , \quad \mathbb{P}\big(I_t = i | I_1, I_2, \dots, I_{t-1}\big) = p_{i,t} \quad a.s.$$
>
> (When $t = 1$, the conditional probability above is just a probability.)
>
> — The statistician incurs the loss $\ell_{I_t,t}$ associated to its choice I_t. He observes $\ell_{I_t,t}$ but does not observe the other losses $\ell_{i,t}$ for $i \neq I_t$.

We now consider the Exp3 algorithm, whose weight vectors $p_t = \big(p_{i,t}\big)_{1 \leqslant i \leqslant K}$ are defined as follows ($\eta > 0$ is a parameter of the algorithm) :

$$p_{i,t} = \frac{\exp\Big(-\eta \sum_{s=1}^{t-1} \widetilde{\ell}_{i,s}\Big)}{\sum_{j=1}^{K} \exp\Big(-\eta \sum_{s=1}^{t-1} \widetilde{\ell}_{j,s}\Big)} \ , \quad 1 \leqslant i \leqslant K \ ,$$

where the estimated losses $\widetilde{\ell}_{i,s}$ are defined by

$$\widetilde{\ell}_{i,s} = \frac{\ell_{i,s}}{p_{i,s}} \, \mathbb{I}_{\{I_s = i\}} \ .$$

(8) Explain why we use $\widetilde{\ell}_{i,s}$ instead of $\ell_{i,s}$.

(9) Show that for all $t \geqslant 1$ and all $i \in \{1, \dots, K\}$,

$$\sum_{t=1}^{T} p_t \cdot \widetilde{\ell}_t \leqslant \min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \widetilde{\ell}_{i,t} + \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{i=1}^{K} p_{i,t} \, \widetilde{\ell}_{i,t}^2 \quad a.s.$$

(10) Define the filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$ by $\mathcal{F}_0 = \{\varnothing, \Omega\}$ (the trivial σ-field) and by $\mathcal{F}_t = \sigma(I_1, \ldots, I_t)$ for all $t \geqslant 1$. Show that $\mathbb{E}\Big[\widetilde{\ell}_{i,t} | \mathcal{F}_{t-1}\Big] = \ell_{i,t}$ almost surely.

(11) Compute $p_t \cdot \widetilde{\ell}_t$ and $\sum_{i=1}^K p_{i,t}\, \widetilde{\ell}_{i,t}^2$, and show that

$$\mathbb{E}\left[\sum_{t=1}^T \ell_{I_t,t} - \min_{1\leqslant i\leqslant K} \sum_{t=1}^T \ell_{i,t}\right] \leqslant \frac{\ln K}{\eta} + \frac{\eta T K}{2}\ .$$

(12) After optimizing in η, which regret bound do we get ? What is the main difference with the regret bound that we derived (during the lectures) for the EWA algorithm ? Why is this bound different ?

Problem 2

Let n be a positive integer, and let $\mathcal{X}$ be a subset of $\mathbb{R}^p$ for some $p > 0$. We assume that there exists a function Φ (called *weak classifier*) which, given any weighted sample $\mathcal{S} = \big\{(x_i, y_i, w_i) : 1 \leqslant i \leqslant n\big\}$, with $x_i \in \mathcal{X}$, $y_i \in \{-1, 1\}$, $0 \leqslant w_i \leqslant 1$ and $w_1 + \cdots + w_n = 1$, yields a classification rule $h = \Psi(\mathcal{S}) : \mathcal{X} \mapsto \{-1, 1\}$ such that

$$\sum_{i=1}^n w_i\, \mathbb{I}\{h(x_i) \neq y_i\} \leqslant \frac{1}{2} - \gamma$$

for some positive real number γ.

Algorithm Adaboost works as follows : for a given number T of iterations,

— **Initialization :** for every $i \in \{1, \ldots, n\}$, let $w_i^1 = 1/n$;

— **Main loop :** for every t from 1 to T :

— compute $h_t = \Phi\big((x_i, y_i, w_i^t)_i\big)$;

— compute

$$\varepsilon_t = \sum_{i=1}^n w_i^t\, \mathbb{I}\{h_t(x_i) \neq y_i\}\ ;$$

— for every $i \in \{1, \ldots, n\}$, let

$$w_i^{t+1} = \frac{w_i^t}{Z_t} \times \exp\big(-\alpha_t\, y_i\, h_t(x_i)\big)\ ,$$

where Z_t is such that $w_1^{t+1} + \cdots + w_n^{t+1} = 1$, and where

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)\ .$$

— **Output :** the final classifier is the function $H : \mathcal{X} \mapsto \{-1, 1\}$ defined by

$$H(x) = \operatorname{sign}\left(\sum_{t=1}^T \alpha_t\, h_t(x)\right)\ ,$$

where $\operatorname{sign}(u) = 2 \times \mathbb{I}\{u \geqslant 0\} - 1$.

We define

$$F(x) = \sum_{t=1}^T \alpha_t\, h_t(x)$$

and

$$e = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{H(x_i) \neq y_i\} \ .$$

(13) In supervised classification, what is the name of e ?

(14) For the case $\mathcal{X} = \mathbb{R}$, propose a possible choice for Φ and give the associated value of γ. If needed, one may assume that the $(x_i)_i$ are all different.

(15) Show that

$$w_i^{T+1} = \frac{\exp\big(-y_i\, F(x_i)\big)}{n\prod_{t=1}^{T} Z_t} \ .$$

(16) Show that

$$e \leqslant \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{y_i\, F(x_i) \leqslant 0\} \ .$$

(17) Show that

$$e \leqslant \frac{1}{n}\sum_{i=1}^{n} \exp\big(-y_i\, F(x_i)\big)$$

(18) Show that

$$\sum_{i=1}^{n} \exp\big(-y_i\, F(x_i)\big) = n\prod_{t=1}^{T} Z_t \ .$$

(19) Show that

$$Z_t = \varepsilon_t \exp(\alpha_t) + (1-\varepsilon_t)\exp(-\alpha_t) \ .$$

(20) What is the value of α that minimizes

$$g(\alpha) = \varepsilon_t \exp(\alpha) + (1-\varepsilon_t)\exp(-\alpha) \ ?$$

(21) Show that

$$Z_t = 2\sqrt{\varepsilon_t(1-\varepsilon_t)} \ .$$

(22) Show that

$$\sum_{i=1}^{n} w_i^{t+1}\, \mathbb{I}\{h_t(x_i) \neq y_i\} = \frac{1}{2} \ .$$

How to interpret this equality ?

(23) For every t between 1 and T, let $\gamma_t = 1/2 - \varepsilon_t$. Show that

$$e \leqslant \prod_{t=1}^{T} \sqrt{1-4\gamma_t^2} \leqslant \exp\left(-2\sum_{t=1}^{T}\gamma_t^2\right) \ .$$

(24) Conclude :

$$e \leqslant \exp\big(-2T\gamma^2\big) \ .$$

(25) Give a value of T_0 such that for every $T \geqslant T_0$, $e = 0$. Explain the sentence : "weak learnability implies strong learnability".

(26) Should one necessarily choose T of the order of T_0 ?

(27) What is the value of T_0 for the value of γ found at Question 14 ?

(28) Why is **Ada**boost said to be *adaptive* ?