CS 7545: Machine Learning Theory **Fall 2018**

Lecture 6: Multiplicative Weight Algorithms

Lecturer: Jacob Abernethy *Scribes: Sebastian Perez-Salazar, Sheng-Tao Yang*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

6.1 Introduction

In the last lecture, we started studying the expert model. In this model, at each time $t = 1, 2 \ldots, T$ we have to decide between two options $\{0, 1\}$ with the aid of N experts. At each time, each of the experts makes a prediction in $\{0, 1\}$ and we have to make a decision $\widehat{y} \in \{0, 1\}$. After this, the nature shows the real value of the option $y \in \{0, 1\}$ and we incur in a loss if $\widehat{y} \neq y$. The goal is to minimize the number of losses. We assumed there exists a perfect expert that never makes mistakes. With this hypothesis, we proved that the halving algorithm achieves $\log N$ mistakes.

In this lecture, we will drop the assumption of the existence of a perfect expert. By a generalization of the halving algorithm, we will construct the *Weighted Majority Algorithm* that will guarantee a number of mistakes no more than 2 times the number of mistakes made by the best expert. Finally, we will introduce the *Exponential Weight Algorithm* that will show tighter bounds.

6.2 Weighted Majority Algorithm

Let $\text{round}(\alpha) = 0$ if $\alpha < \frac{1}{2}$ and 1 if $\alpha \geq \frac{1}{2}$. We present the Weighted Majority Algorithm:

Algorithm 1: Weighted Majority Algorithm (WMA)

Parameter $\varepsilon \in [0, 1]$ N experts making predictions.
We maintain weights $w_i^t \in [0, 1]$ for each expert $i = 1, \ldots, N$. Initially, $w_i^1 = 1$ for all i.
for $t = 1, \ldots, T$ **do**
  Expert i predicts $x_i^t \in \{0, 1\}$, $i = 1, \ldots, N$.
  Predict $\widehat{y} = \text{round}\left(\frac{\sum_{i=1}^N w_i^t x_i^t}{\sum_{i=1}^N w_i^t}\right)$. Observe $y^t \in \{0, 1\}$.
  Update $w_i^{t+1} = w_i^t (1 - \varepsilon)^{\mathbf{1}[x_i^t \neq y^t]}$.

Observe that halving algorithm is WMA with $\varepsilon = 1$.

Definition 6.1 *We define the number of mistakes made by WMA as* $M_T(WMA) = \sum_{t=1}^T \mathbf{1}[\widehat{y}^t \neq y^t]$. *In the same way, we define the the number of mistakes made by expert* i *as* $M_T(i) = \sum_{t=1}^T \mathbf{1}[x_i^t \neq y^t]$.

Theorem 6.2 *For any expert* i *and* $\varepsilon \in (0, \frac{1}{2})$, *WMA guarantees* $M_T(WMA) \leq 2\frac{\log N}{\varepsilon} + 2(1 + \varepsilon) M_T(i)$.

Before proving the theorem, we will introduce some useful bounds for exponential function and logarithm function.

1. For any $x \in (-1, +\infty)$, $\log(1 + x) \leq x$. The proof follows by the concavity of $\log(1 + x)$.
2. $1 + x \leq e^x$ for any x. For $x \leq -1$ is clearly true, and for $x > -1$ we can use the monotonicity of exp and the previous inequality.

3. $e^{\alpha x} \leq 1 + (e^{\alpha} - 1)x$ for any $x \in [0, 1]$. Follows by using the convexity of $e^{\alpha x}$ in x.
4. $-\log(1 + x) \leq -x + x^2$ for any $x \in [0, \frac{1}{2}]$. **[Challenge]**.

Now, we are ready to present the proof of Theorem 6.2.
Proof: We will use a potential argument. Let $\Phi_t = \sum_{i=1}^{N} w_i^t$. We point three facts about this potential:

1. $\Phi_1 = N$.
2. $\Phi_{T+1} \geq w_i^{T+1} = (1 - \varepsilon)^{M_T(i)}$ for any expert $i = 1, \ldots, N$.
3. $\Phi_{T+1} \leq \Phi_1(1 - \frac{\varepsilon}{2})^{M_T(\text{WMA})}$. For this, clearly we have $\Phi_{t+1} \leq \Phi_t$. Moreover, if WMA makes a mistake at time t, then the overall weight of experts making a mistake is at least half of the total weight Φ_t. More formally, we write
$$\Phi_t = \beta + (\Phi_t - \beta),$$
where β is the total weight of incorrect experts. Since the algorithm made a mistake, we have $\beta \geq \frac{\Phi_t}{2}$. Then, since all wrong experts are penalized by $(1 - \varepsilon)$, we obtain
$$\Phi_{t+1} = \beta(1 - \varepsilon) + (\Phi_t - \beta) = \Phi_t - \beta\varepsilon \leq \Phi_t \left(1 - \frac{\varepsilon}{2}\right).$$

Now, by using the three facts we obtain

$$(1 - \varepsilon)^{M_T(i)} \leq \Phi_{T+1} \leq N \left(1 - \frac{\varepsilon}{2}\right)^{M_T(\text{WMA})}.$$

By taking $-\log(\cdot)$ and using inequalities 1 and 4 we obtain

$$M_T(i)(\varepsilon + \varepsilon^2) \geq -M_T(i)\log(1 - \varepsilon) \geq -\log N - M_T(\text{WMA})\log\left(1 - \frac{\varepsilon}{2}\right) \geq -\log N + M_T(\text{WMA})\frac{\varepsilon}{2},$$

or equivalently $M_T(\text{WMA}) \leq 2\frac{\log N}{\varepsilon} + 2(1 + \varepsilon)M_T(i)$. ■

Observations

- If we take minimum in i, the theorem states that the overall mistake made by the algorithm is roughly at most 2 times the mistakes made by the best expert.
- The factor 2 appears because of the rounding used in the algorithm. In the next section, we exhibit an algorithm that overcomes this issue.

6.2.1 Exponential Weighted Algorithm

In Weighted Majority Algorithm, we used the number of mistakes as a benchmark. In order to generalize the previous algorithm, we first assume the algorithm can make predictions with values between 0 and 1. That is,

$$\widehat{y} = \frac{\sum_{i=1}^{N} w_i^t x_i^t}{\sum_{i=1}^{N} w_i^t}.$$

We also generalize the benchmark of the algorithm. A *convex loss function* is a function $\ell : [0, 1] \times \{0, 1\} \to \mathbf{R}$ such that ℓ is a convex function in its first argument. For example,

- Square loss: $\ell(\hat{y}, y^t) = (\hat{y} - y^t)^2$.
- Absolute loss: $\ell(\hat{y}, y^t) = |\hat{y} - y^t|$.
- Logarithm loss: $\ell(\hat{y}, y^t) = y^t \log\hat{y} + (1 - y^t)\log(1 - \hat{y})$.

Further, for a specific expert, we let its weight w_i^t decays exponentially according to the loss in each round. As a result, we propose the Exponential Weights Algorithm:

Algorithm 2: Exponential Weights Algorithm (EWA)

Parameter $\eta > 0$, N experts making predictions.
We maintain weights $w_i^t \in [0,1]$ for each expert $i = 1, \ldots, N$. Initially, $w_i^1 = 1$ for all i.
for $t = 1, \ldots, T$ **do**
    Expert i predicts $x_i^t \in \{0,1\}$, $i = 1, \ldots, N$.
    Predict $\widehat{y} = \frac{\sum_{i=1}^N w_i^t x_i^t}{\sum_{i=1}^N w_i^t}$. Observe $y^t \in \{0,1\}$.
    Update $w_i^{t+1} = w_i^t \exp(-\eta \ell(x_i^t, y^t))$.

To measure the performance of EWA, we introduce the notion of the algorithm loss and the expert loss. This helps us to define the notion of regret from round 1 to round T.

Definition 6.3 *We define the loss function of EWA as* $L_T(EWA) = \sum_{t=1}^T \ell(\hat{y}^t, y^t)$. *Similarly, we define the loss function of expert* i *as* $L_T(i) = \sum_{t=1}^T \ell(x_i^t, y^t)$

In words, the loss function measures the difference between predictions and true outcomes with respect to a fixed loss function ℓ.

Definition 6.4 *We define the regret of EWA as*

$$\begin{aligned} Regret &= L_T(EWA) - \min_{1 \le i \le N} L_T(i) \\ &= \sum_{t=1}^T \ell(\hat{y}^t, y^t) - \min_{1 \le i \le N} \sum_{t=1}^T \ell(x_i^t, y^t) \end{aligned} \tag{6.1}$$

Intuitively, this value measures the regret of being in an online setting as opposed to be in an offline full-information setting.

Theorem 6.5 *(Upper bound of* $L_T(EWA)$*) The EWA satisfies*

$$L_T(EWA) \le \frac{\eta L_T(i) + \log N}{1 - \exp(-\eta)}.$$

Corollary 6.6 *(Upper bound of Regret) By tunning* λ *appropriately, we have*

$$L_T(EWA) - L_T(i^*) \le \log N + \sqrt{2 L_T(i^*) \mathrm{N}},$$

where $i^* = \mathrm{argmin} L_T(i)$ *is the index of the best expert.*

Observe that the left hand side corresponds to the regret. Moreover, if the loss function is bounded, say by 1, then certainly $L_T(i^*)$ is less than T. Therefore, $\frac{Regret}{T} \le o(\sqrt{T})$, that is, in average, the loss of EWA approaches to the loss made by the best expert.

Lemma 6.7 *Let* X *be any random variable in* $[0,1]$ *and* $s \in \mathbb{R}$*, then*

$$\log \mathbb{E} e^{sX} \le (e^s - 1) \mathbb{E} X$$

Proof: Since $e^{sx} \le 1 + (e^s - 1)x \ \forall x \in [0,1]$, by taking expectation on both sides, we have

$$\mathbb{E} e^{sX} \le 1 + (e^s - 1) \mathbb{E} X.$$

Thus,

$$\log \mathbb{E} e^{sX} \le \log(1 + (e^s - 1)\mathbb{E} X) \le (e^s - 1)\mathbb{E} X.$$

■