Machine Learning Theory (CSC 482A/581A) - Lectures 28 & 29

Nishant Mehta

1 Prediction with Expert Advice

We now upgrade the decision-theoretic online learning setting to a more general setting known as prediction with expert advice. In this setting, we have a loss function $\ell \colon \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ that, for each action a in an action space $\mathcal{A}$ and each outcome y in an outcome space $\mathcal{Y}$, produces a loss $\ell(a, y)$. We will assume that the action space $\mathcal{A}$ is convex and that the loss function is convex as a function of its first argument (the action $a \in \mathcal{A}$). Two common examples are:

- Classification with absolute loss: Here, we take $\mathcal{A} = [0, 1]$, $\mathcal{Y} = \{0, 1\}$, and $\ell(a, y) = |a - y|$.
- Classification with squared loss: We take $\mathcal{A}$ and $\mathcal{Y}$ as before and now set $\ell(a, y) = (a - y)^2$.

In prediction with expert advice, each of the K experts now provides advice in the form of a suggested action from $\mathcal{A}$ at the start of each round. Learner then aggregates these actions in some way, producing its own action within $\mathcal{A}$. Finally, Nature selects an outcome, and Learner and each expert suffer loss according to their respective actions and the outcome.

Formally, the protocol is as follows:

Protocol:

For round $t = 1, 2, \ldots$

1. Nature selects the expert advice $\{f_{j,t} : j \in [K]\}$ and reveals it to Learner.
2. Learner selects action $a_t \in \mathcal{A}$.
3. Nature selects an outcome $y_t \in \mathcal{Y}$ and reveals it to Learner.
4. Each expert $j \in [K]$ suffers loss $\ell(f_{j,t}, y_t)$ and Learner suffers loss $\ell(a_t, y_t)$.

As before, our goal is to minimize the regret, now defined as:

$$\sum_{t=1}^{T} \ell(a_t, y_t) - \min_{j \in [K]} \sum_{t=1}^{T} \ell(f_{j,t}, y_t).$$

To simplify the presentation, we will adopt the following notation for any $j \in [K]$ and $t \in [T]$:

- $\ell_{j,t} = \ell(f_{j,t}, y_t)$;
- $L_{j,t} = \sum_{s=1}^{t} \ell_{j,s}$.

Also, for any $t \in [T]$, denote the loss and cumulative loss of the learning algorithm as

- $\hat{\ell}_t = \ell(a_t, y_t)$;

- $\hat{L}_t = \sum_{s=1}^t \hat{\ell}_s$.

The algorithm that we study for this setting is a suitably adapted variation of the exponential weights algorithm. This algorithm, called the exponentially weighted average forecaster, works as follows. In each round, the algorithms maintains weights over the experts, with $w_{j,t}$ indicating the weight on the j^{th} expert in round t. In round t, the forecaster predicts according to the following weighted average of the experts' actions:

$$a_t = \frac{\sum_{j=1}^K w_{j,t-1} f_{j,t}}{\sum_{j=1}^K w_{j,t-1}}.$$

The weights are initialized as $w_{j,0} = 1$ for $j \in [K]$. At the end of a given round, the losses of the experts' are observable, and the weights are updated according to the rule

$$w_{j,t} = w_{j,t-1} e^{-\eta \ell_{j,t}}.$$

By unrolling this update backwards to $w_{j,0}$, we see that

$$w_{j,t} = e^{-\eta L_{j,t}}.$$

From the above, we can see that the weight updates precisely match the updates in Hedge.

Moreover, since we have assumed that the loss is convex, a nearly identical analysis as we used for Hedge implies the following worst-case regret guarantee.

Theorem 1. *Let the learning rate η be set as $\eta = \sqrt{\frac{8 \log K}{T}}$. Then, for any sequence of expert predictions $(f_{j,t})_{j \in [K], t \in [T]}$ and any sequence of outcomes $y_1, \ldots, y_T$, the regret of the exponentially weighted average forecaster satisfies*

$$\hat{L}_T - \min_{j \in [K]} L_{j,T} \leq \sqrt{\frac{T \log K}{2}}.$$

Proof. The proof of this result requires only a minor modification to the proof of Theorem 1 from Lecture 25. We recall the 3 steps of that proof and indicate where the analysis needs to be adapted.

For $t \in [T]$, define

$$W_t := \sum_{j=1}^K w_{j,t}.$$

The first step is to show that

$$\log \frac{W_T}{W_0} \geq -\eta \min_{j \in [K]} L_{j,T} - \log K. \tag{1}$$

The analysis for this step, as already done for Hedge, holds without modification.

The second step is to show that for any $t \in [T]$,

$$\log \frac{W_t}{W_{t-1}} \leq -\eta \, \mathsf{E}_{j \sim p_t}[\ell_{j,t}] + \frac{\eta^2}{8}, \tag{2}$$

where p_t is the distribution over $[K]$ played by Hedge in round t. This distribution is defined as

$$p_t(j) = \frac{w_{j,t-1}}{W_{t-1}}.$$

The claim (and proof) for this step from Hedge needs to be adapted, since $\mathsf{E}_{j \sim p_t}[\ell_{j,t}]$ is the loss of Hedge in round t, but it is not the loss of the exponentially weighted average forecaster in round t. Since the loss $\ell_{j,t} = \ell(f_{j,t}, y_t)$ is convex in its first argument, Jensen's inequality implies that

$$\mathsf{E}_{j \sim p_t}[\ell(f_{j,t}, y_t)] \geq \ell(\mathsf{E}_{j \sim p_t}[f_{j,t}], y_t) = \hat{\ell}_t,$$

which, combined with (2), implies that

$$\log \frac{W_t}{W_{t-1}} \leq -\eta \hat{\ell}_t + \frac{\eta^2}{8}. \tag{3}$$

The remainder of the proof of Theorem 1 from Lecture 25 can be retraced to yield the result, where we sum (3) from $t = 1$ to T, combine the resulting inequality with (1), and use the specified setting of η. $\square$

2 Exp-concave losses

We thus have have seen regret bounds that scale as $\sqrt{T \log K}$. We now turn to a special types of loss functions, known as an exp-concave losses. These loss functions are of interest because

- They encompass several well-known and widely-used loss functions, including squared loss, logistic loss, and log loss.
- Remarkably, for these loss functions the exponentially weighted average forecaster achieves regret that is *constant* with respect to T.

Definition 1. We say that a loss function ℓ is η-exp-concave if, for each outcome $y \in \mathcal{Y}$, the function $a \mapsto e^{-\eta \ell(a,y)}$ is concave. Equivalently, ℓ is η-exp-concave if, for all $y \in \mathcal{Y}$ and all distributions P over $\mathcal{A}$,

$$\mathsf{E}_{a \sim P}\left[e^{-\eta \ell(a,y)}\right] \leq e^{-\eta \ell(\mathsf{E}_{a \sim P}[a], y)}. \tag{4}$$

Before showing how to get an improved regret bound for exp-concave losses, let's first take a look at a few examples.

The first and simplest example is log loss. Prediction with expert advice with log loss is specified by taking $\mathcal{A} = [0, 1]$, $\mathcal{Y} = \{0, 1\}$, and $\ell(a, y) = -y \log a - (1-y) \log(1-a)$. Log loss is 1-exp-concave, as is readily verified by considering the two cases. For instance, if $y = 1$, then the function

$$a \mapsto e^{-\ell(a,1)} = e^{\log a} = a$$

is clearly concave.

In class I also gave an example with sequential investment and a variant of log loss - I will try to add this to the notes soon

The second example is squared loss, with $\mathcal{A} = \mathcal{Y} = [0, 1]$ and $\ell(a, y) = (a - y)^2$. In order to establish the exp-concavity of squared loss, we will use an alternate characterization of exp-concavity. For the time being, we restrict to one-dimensional actions a for simplicity. Take $\mathcal{X} \subset \mathbb{R}$; recall that a function $g \colon \mathcal{X} \to \mathbb{R}$ is concave if $g''(x) \leq 0$ for all $x \in \mathbb{R}$. Now, using the definition of η-exp-concavity, we see that a function $f \colon \mathcal{X} \to \mathbb{R}$ is η-exp-concave if and only if

$$\eta^2 (f'(x))^2 e^{-\eta f(x)} - \eta f'(x) e^{-\eta f(x)} \leq 0 \qquad \text{for all } x \in \mathcal{X},$$

which is equivalent to the condition

$$\eta(f'(x))^2 \le f''(x) \qquad \text{for all } x \in \mathcal{X}.$$

Returning to the example of squared loss, we can verify that the squared loss is η-exp-concave if and only if

$$\eta\,(2(a-y))^2 \le 2 \qquad \text{for all } a, y \in [0,1],$$

or equivalently,

$$(a-y)^2 \le \frac{1}{2\eta} \qquad \text{for all } a, y \in [0,1].$$

This condition is satisfied for $\eta = \frac{1}{2}$, and so the squared loss is $\frac{1}{2}$-exp-concave.

3 Constant regret under exp-concavity

Theorem 2. *Let $\ell \colon \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ be an η-exp-concave loss for some $\eta > 0$. Let the learning rate be set to the same value η. Then, for any sequence of expert predictions $(f_{j,t})_{j \in [K], t \in [T]}$ and any sequence of outcomes $y_1, \ldots, y_T$, the regret of the exponentially weighted average forecaster satisfies*

$$\hat{L}_T - \min_{j \in [K]} L_{j,T} \le \frac{\log K}{\eta}.$$

Proof. The proof of this result is remarkably simpler than the proof of Theorem 1. First, observe that the regret satisfies

$$\begin{aligned}
\hat{L}_T - \min_{j \in [K]} L_{j,T} &= \max_{j \in [K]} \left\{ \hat{L}_T - L_{j,T} \right\} \\
&= \frac{1}{\eta} \log \max_{j \in [K]} e^{\eta(\hat{L}_T - L_{j,T})} \\
&\le \frac{1}{\eta} \log \sum_{j \in [K]} e^{\eta(\hat{L}_T - L_{j,T})} \\
&= \Phi(T),
\end{aligned}$$

where, for each $t \in [T]$, we define the potential function

$$\Phi(t) = \frac{1}{\eta} \log \sum_{j \in [K]} e^{\eta(\hat{L}_t - L_{j,t})}.$$

Next, we claim that, for any $t \in [T]$,

$$\Phi(t) \le \Phi(t-1) \tag{5}$$

We will prove this claim momentarily. Supposing for now that the claim is true, then

$$\begin{aligned}\hat{L}_T - \min_{j\in[K]} L_{j,T} \leq \frac{\log K}{\eta} &\leq \Phi(T) \\ &\leq \Phi(T-1) \\ &\dots \\ &\leq \Phi(0) \\ &= \frac{1}{\eta}\log \sum_{j\in[K]} e^{\eta(\hat{L}_0 - L_{j,0})} \\ &= \frac{1}{\eta}\log \sum_{j\in[K]} e^{\eta 0} \\ &= \frac{\log K}{\eta},\end{aligned}$$

and so the result follows.

Finally, we prove (5). Observe that it is equivalent to prove that

$$\sum_{j\in[K]} e^{\eta(\hat{L}_t - L_{j,t})} \leq \sum_{j\in[K]} e^{\eta(\hat{L}_{t-1} - L_{j,t-1})},$$

which itself is equivalent to proving that

$$\sum_{j\in[K]} e^{-\eta L_{j,t-1}} e^{-\eta \ell_{j,t}} e^{\eta \hat{\ell}_t} \leq \sum_{j\in[K]} e^{-\eta L_{j,t-1}}.$$

Now, using $w_{j,t-1} = e^{-\eta L_{j,t-1}}$ and rearranging, this is equivalent to

$$\frac{\sum_{j\in[K]} w_{j,t-1} e^{-\eta \ell_{j,t}}}{\sum_{j\in[K]} w_{j,t-1}} \leq e^{-\eta \hat{\ell}_t}. \tag{6}$$

Finally, setting $p_{j,t} = \frac{w_{j,t-1}}{\sum_{i=1}^K w_{i,t-1}}$ and recalling that

$$\ell_{j,t} = \ell(f_{j,t}, y_t) \qquad \hat{\ell}_t = \ell(a_t, y_t) = \ell\left(\sum_{j=1}^T p_{j,t} f_{j,t}, y_t\right),$$

Thus, (6) becomes

$$\mathsf{E}_{j\sim p_t}\left[e^{-\eta \ell(f_{j,t}, y_t)}\right] \leq e^{-\eta \ell(\mathsf{E}_{j\sim p_t}[f_{j,t}], y_t)}.$$

This last inequality holds because of ℓ is η-exp-concave. $\square$