Complementi di Algoritmi e Strutture Dati

Exponential weights: when experts become bandits

Written by *Tommaso R. Cesari* versione 8 giugno 2018

Consider an online decision-making problem in which an algorithm has to respond to a sequence of requests that arrive one by one. The algorithm must take an action as each request arrives, and it may discover later that its past actions were suboptimal. However, once they are played, past actions cannot be changed. We begin by illustrating this setting with a concrete example.

Imagine the process of picking good times to invest in a stock. For simplicity, assume that there is a single stock of interest and its daily price movement is modeled as a sequence of binary events: *up* or *down*. (This will be generalized later to allow non-binary events.) Each morning we try to predict whether the price will go up or down that day; if our prediction happens to be wrong we lose a dollar that day, and if it is correct, we lose nothing. We let the stock movements in the model be arbitrary and even adversarial, meaning that there might be an antagonistic environment which picks the movements with the deliberate purpose of causing us the largest possible loss. To balance out this pessimistic assumption, we assume that while making our predictions, we are allowed access to the predictions of K "experts". These experts could be arbitrarily correlated, and they may or may not be able to make reliable predictions. The algorithm's goal is to limit its cumulative losses (i.e., its bad predictions) to roughly the same as the best of these experts.

At first sight this seems an impossible goal, since it is not known until the end of the sequence who the best expert was, whereas the algorithm is required to make predictions along the way. For example, a first, naive algorithm may compute each day's up or down prediction by going with the majority opinion among the experts that day. However, it is easy to see that this algorithm is doomed to failure because a majority of experts may be consistently wrong on every single day. A better way to pick a prediction consists in maintaining a weighting of the experts. Initially all have equal weight. As time goes on, some experts are seen as making better predictions than others, and the algorithm increases their weight proportionally. The algorithm's prediction of up or down for each day is then computed by going with the opinion of the *weighted* majority of the experts for that day. This way, experts that made good predictions in the past are preferred to ones who performed poorly. A famous implementation of this concept is an algorithm called Hedge. Before going through its mechanics in details we introduce rigorously the aforementioned online decision-making problem, called *prediction from expert advice*.

Prediction with expert advice is based on the following protocol for sequential decisions. A finite set of experts $\{1, \dots, K\}$ is fixed and known by both the decision maker and the (adversarial) environment. At each round $t = 1, 2, \dots$ the environment secretly chooses a loss $\ell_t(i) \in [0, 1]$ for each expert i; the decision maker picks an expert I_t (possibly at random), then he or she incurs a loss $\ell_t(I_t)$ and the losses $\ell_t(i)$ of all experts i are revealed. The performance of the decision maker at time T is measured by the difference between its sequential risk and the average loss of the best expert for $\ell_1, \dots, \ell_T$, that is

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\ell_t(I_t)\right] - \min_{i=1,\dots,K}\left(\frac{1}{T}\sum_{t=1}^{T}\ell_t(i)\right)$$

where the expectation is taken with respect to the draw of $I_1, \ldots, I_T$.

Algorithm Hedge
Parameter: learning rate $\gamma \in (0,1)$
Initialization: $w_1(i)$ for all $i = 1, \ldots, K$
For $t = 1, \ldots, T$

1. define the distribution $p_t(i) = w_t(i)/W_t$, where $W_t = \sum_{j=1}^K w_t(j)$
2. draw I_t according to p_t
3. incur a loss $\ell_t(I_t)$ and $\ell_t(i)$ is revealed for each expert i
4. update the weights
$$w_{t+1}(i) = w_t(i)e^{-\gamma \ell_t(i)}$$

We want to prove now that if $I_1, \ldots, I_T$ are picked by Hedge and $T \to \infty$ then

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \ell_t(I_t)\right] - \min_{i=1,\ldots,K}\left(\frac{1}{T}\sum_{t=1}^T \ell_t(i)\right) \to 0$$

i.e., that the sequential risk of Hedge converges to the average loss of the optimal expert.

The analysis looks at the ratio between the total weight on the experts at subsequent rounds:

$$\begin{aligned}\frac{W_{t+1}}{W_t} &= \sum_{i=1}^K \frac{w_{t+1}(i)}{W_t} = \sum_{i=1}^K \frac{w_t(i)e^{-\gamma\ell_t(i)}}{W_t} = \sum_{i=1}^K p_t(i)e^{-\gamma\ell_t(i)} \\ &\le \sum_{i=1}^K p_t(i)\Big(1 - \gamma\ell_t(i) + \gamma^2\ell_t(i)^2/2\Big) = 1 - \gamma\sum_{i=1}^K p_t(i)\ell_t(i) + \frac{\gamma^2}{2}\sum_{i=1}^K p_t(i)\ell_t(i)^2\end{aligned}$$

where the inequality follows by $e^{-x} \le 1 - x + x^2/2$ (which holds for all $x \ge 0$). Taking logs, we obtain

$$\ln\left(\frac{W_{t+1}}{W_t}\right) \le \ln\left(1 - \gamma\sum_{i=1}^K p_t(i)\ell_t(i) + \frac{\gamma^2}{2}\sum_{i=1}^K p_t(i)\ell_t(i)^2\right)$$

Using $\ln(1+x) \le x$ (which holds for all $x > -1$) and summing over $1, \ldots, T$ yields

$$\ln\left(\frac{W_{T+1}}{W_1}\right) = \sum_{t=1}^T \ln\left(\frac{W_{t+1}}{W_t}\right) \le -\gamma\sum_{t=1}^T\sum_{i=1}^K p_t(i)\ell_t(i) + \frac{\gamma^2}{2}\sum_{t=1}^T\sum_{i=1}^K p_t(i)\ell_t(i)^2$$

On the other hand, for any fixed arm k

$$\ln\left(\frac{W_{T+1}}{W_1}\right) \ge \ln\left(\frac{w_{T+1}(k)}{W_1}\right) = -\gamma\sum_{t=1}^T \ell_t(k) - \ln(K)$$

Putting together the two bounds on $\ln(W_{T+1}/W_1)$ and dividing by γ gives

$$\sum_{t=1}^T\sum_{i=1}^K p_t(i)\ell_t(i) - \sum_{t=1}^T \ell_t(k) \le \frac{\ln K}{\gamma} + \frac{\gamma}{2}\sum_{t=1}^T\sum_{i=1}^K p_t(i)\ell_t(i)^2 \tag{1}$$

Note now that

$$\mathbb{E}[\ell_t(I_t)] = \sum_{i=1}^{K} p_t(i)\ell_t(i) \tag{2}$$

by definition of expectation. Furthermore, upper bounding $\ell_t(i)^2$ by 1 and using the fact that p_t is a distribution we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \sum_{t=1}^{T} \ell_t(k) \leq \frac{\ln K}{\gamma} + \frac{\gamma}{2}T$$

Being the previous inequality true for all experts k and all learning rates γ, tuning $\gamma = \sqrt{2\ln(K)/T}$ and dividing both sides by T yields

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{i=1,\dots,K}\left(\frac{1}{T}\sum_{t=1}^{T} \ell_t(i)\right) \leq \sqrt{\frac{2\ln K}{T}}$$

This proves that the sequential risk of Hedge converges to the average loss of the optimal expert.

Surprisingly, a similar result can be achieved even if the only loss that is revealed at each round is $\ell_t(I_t)$. Imagine the problem of placing ads on the Web. For each incoming user $t = 1, 2, \dots$ a publisher selects an ad I_t from a pool of K ads, and displays the corresponding ad to the user. The publisher then loses 1 if the ad is not clicked by the user, otherwise he or she loses 0. After each interaction the publisher finds out if the user did or did not click on the chosen ad but he or she has no way of knowing if the user would have clicked on any of the other ads.

This problem is well-modeled by the following protocol for sequential decisions, called *multi-armed bandit* setting. A finite set of actions $\{1, \dots, K\}$ is fixed and known by both the decision maker and the (adversarial) environment. At each round $t = 1, 2, \dots$ the environment secretly chooses a loss $\ell_t(i) \in [0, 1]$ for each action i; the decision maker picks an action I_t (possibly at random), then he or she incurs a loss $\ell_t(I_t)$ and only the loss of the chosen action is revealed. The performance of the decision maker at time T is measured by the difference between its sequential risk and the average loss of the best action for $\ell_1, \dots, \ell_T$, that is the difference

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{i=1,\dots,K}\left(\frac{1}{T}\sum_{t=1}^{T} \ell_t(i)\right)$$

where the expectation is taken with respect to the draw of $I_1, \dots, I_T$.

It is possible to modify Hedge in such a way that the rate of convergence $\mathcal{O}(T^{-1/2})$ is preserved. The idea is substituting $\ell_t(i)$ (which is never revealed) in the update of the weights (line 4) with an unbiased estimate of it. This is done by a famous algorithm called Exp3 using "importance-weighted" estimators

$$\widehat{\ell}_t(i) = \frac{\ell_t(i)}{p_t(i)}\mathbb{I}\{I_t = i\}$$

Being $\widehat{\ell}_t(i)$ a random variable, $p_t(i)$ is also a random variable. However $p_t(i)$ is constant for any given realization of $I_1, \dots, I_{t-1}$. Hence

$$\mathbb{E}\left[\widehat{\ell}_t(i) \mid I_1, \dots, I_{t-1}\right] = \ell_t(i) \qquad \text{and} \qquad \mathbb{E}\left[\widehat{\ell}_t(i)^2 \mid I_1, \dots, I_{t-1}\right] = \frac{\ell_t(i)^2}{p_t(i)} \leq \frac{1}{p_t(i)}$$

Proceeding as in the analysis of Hedge, equation (1) can be derived with $\widehat{\ell}_t$ instead of ℓ_t. Exploiting the previous equalities we obtain, for any action k and any learning rate γ,

$$\sum_{t=1}^{T}\sum_{i=1}^{K} p_t(i)\ell_t(i) - \sum_{t=1}^{T}\ell_t(k) \leq \frac{\ln K}{\gamma} + \frac{\gamma}{2}KT \tag{3}$$

Similarly to equation (2),

$$\mathbb{E}[\ell_t(I_t) \mid I_1, \ldots, I_{t-1}] = \sum_{i=1}^{K} p_t(i)\ell_t(i)$$

By the tower rule of the expectation this implies

$$\mathbb{E}[\ell_t(I_t)] = \mathbb{E}\left[\sum_{i=1}^{K} p_t(i)\ell_t(i)\right] \tag{4}$$

Putting (3) and (4) together, dividing by T and picking $\gamma = \sqrt{2\ln(K)/(KT)}$ yields

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\ell_t(I_t)\right] - \min_{i=1,\ldots,K}\left(\frac{1}{T}\sum_{t=1}^{T}\ell_t(i)\right) \leq \sqrt{\frac{2K\ln K}{T}}$$

This is quite surprising as, up to a $\sqrt{K}$ factor, this is exactly the same rate of convergence achievable in the *prediction with expert advice* setting where *all* losses are revealed after each round. The extra $\sqrt{K}$ term can be seen as a consequence of the fact that in the *multi-armed bandit* setting at each round we only see $1/K$-th of the total number of losses.