Online Learning

Pierre Gaillard

November 2018

1 Introduction

In many applications, the data set is not available from the beginning to learn a model but it is observed sequentially as a flow of data. Furthermore, the environment may be so complex that it is unfeasible to choose a comprehensive model and use classical statistical theory and optimization. A classic example is the spam detection which can be seen as a game between spammer and spam filters. Each trying to fool the other one. Another example, is the prediction of processes that depend on human behaviors such as the electricity consumption. These problems are often not adversarial games but cannot be modeled easily and are surely not i.i.d.
There is a necessity to take a robust approach by using a method that learns as ones goes along, learning from experiences as more aspects of the data and the problem are observed. This is the goal of online learning. The curious reader can know more about online learning in the books Cesa-Bianchi and Lugosi [2006], Hazan et al. [2016], Shalev-Shwartz et al. [2012].

Setting In online learning, a player sequentially makes decisions based on past observations. After committing the decision, the player suffers a loss (or receives a reward depending on the problem). Every possible decision incurs a (possibly different) loss. The losses are unknown to the player beforehand an may be arbitrarily chosen by some adversary. More formally, an online learning problem can be formalized as follows.

At each time step $t = 1, \dots, T$
- the player chooses an action $x_t \in \mathcal{X}$ (compact decision set; $\mathcal{X} = \{1, \dots, K\}$ in this lecture);
- the environment chooses a loss function $\ell_t : \mathcal{X} \to [0, 1]$;
- the player suffers loss $\ell_t(x_t)$ and observes
    - the losses of every actions: $\ell_t(x)$ for all $x \in \mathcal{X}$ → full-information feedback
    - the loss of the chosen action only: $\ell_t(x_t)$ → bandit feedback.

The goal of the player is to minimize his cumulative loss:

$$\widehat{L}_T = \sum_{t=1}^{T} \ell_t(x_t)\,.$$

Of course, if the environment chooses large losses $\ell_t(x)$ for all decisions $x \in \mathcal{X}$, it is impossible for the player to ensure small cumulative loss. Therefore, one needs a relative criterion: the regret of the player is the difference between the cumulative loss he incurred and that of the best fixed decision in hindsight.

Definition 1 (Regret)**.** *The regret of the player after T time steps is*

$$R_T := \sum_{t=1}^{T} \ell_t(x_t) - \min_{x \in \mathcal{X}} \sum_{t=1}^{T} \ell_t(x)\,.$$

The goal of the player is to ensure a sublinear regret $R_T = o(T)$ as $T \to \infty$ and this for any possible sequence of losses $\ell_1, \dots, \ell_T$. In this case, the average performance of the player will approach on the long term the one of the best decision.

2 Full information feedback

We will start with the simple case of finite decision set $\mathcal{X} = \{1, \dots, K\}$ with full information feedback. At each time $t \geq 1$, the player chooses $x_t \in \{1, \dots, K\}$, the environment chooses a loss vector $\ell_t = (\ell_t(1), \dots, \ell_t(K)) \in [0,1]^K$ corresponding to the loss of each action and the player observes ℓ_t. This particular case is often considered as *prediction with expert advice*: each action corresponding to some expert advice.

Need of a random strategy The following proposition shows that the choice x_t cannot be deterministic. Otherwise the adversary may fool the player by taking ℓ_t depending on x_t.

Proposition 1. *Any deterministic algorithm may incur a linear regret. In other words, we can find some sequence of losses ℓ_t such that $R_T \gtrsim T$.*

Proof. Since x_t is deterministic, the loss function ℓ_t can depend on x_t. We then choose $\ell_t(x_t) = 1$ and $\ell_t(x) = 0$ for $x \neq x_t$. Then one of the chosen actions was picked less then T/K times so that $\max_{1 \leq k \leq K} \ell_t(k) \leq T/K$. Therefore, $R_T \geq (1 - 1/K)T$. □

From the above proposition, we see that the strategy of the learner needs to be random. Therefore, instead of choosing an action in $\{1, \dots, K\}$, the player chooses a probability distribution $p_t \in \Delta_K := \{p \in [0,1]^K : \sum_k p_k = 1\}$ and draws $x_t \sim p_t$.

The regret R_T will thus be a random quantity that depends on the randomness of the algorithm (and eventually of the data). We will thus focus on upper-bounding the regret:

- with high-probability: $R_T \leq \varepsilon$ with probability at least $1 - \delta$;
- in expectation: $\mathbb{E}[R_T] \leq \varepsilon$.

Note that since the losses are bounded in $[0,1]$ a bound in high probability entails a bound in expectation. If $R_T \leq \varepsilon$ with probability at least $1 - \delta$, then

$$\begin{aligned} \mathbb{E}[R_T] &\leq \mathbb{E}\big[R_T | R_T \leq \varepsilon\big]\mathbb{P}(R_T \leq \varepsilon) + \mathbb{E}\big[R_T | R_T \geq \varepsilon\big]\mathbb{P}(R_T \geq \varepsilon) \\ &\leq \varepsilon + T\delta \,. \end{aligned} \tag{1}$$

How to choose the weights p_t? The idea is to give more weight to actions that performed well in the past. But we should not give all the weight to the current best action, otherwise the algorithm is deterministic and the regret may be linear. The exponentially weighted average forecaster (EWA) also called Hedge performs this trade-off by choosing a weight that decreases exponentially fast with the past errors.

```
The Exponentially weighted average forecaster (EWA)
Parameter: $\eta > 0$
Initialize: $p_1 = \left(\frac{1}{K}, \dots, \frac{1}{K}\right)$
For $t = 1, \dots, T$
    – draw $x_t \sim p_t$; incur loss $\ell_t(x_t)$ and observe $\ell_t \in [0,1]^K$;
    – update for all $k \in \{1, \dots, K\}$
          $p_{t+1}(k) = \frac{e^{-\eta \sum_{s=1}^{t} \ell_s(k)}}{\sum_{j=1}^{K} e^{-\eta \sum_{s=1}^{t} \ell_s(j)}}$ .
```

The following theorem proves that EWA achieves sublinear regret.

Theorem 1. *Let $T \geq 1$. For all sequence of losses $\ell_1, \dots, \ell_T \in [0,1]^K$, EWA achieves almost surely for the choice $\eta = \sqrt{\frac{\log K}{T}}$ the regret bound*

$$\mathbb{E}[R_T] \leq 2\sqrt{T \log K}\,.$$

Proof. We denote $W_t(j) = e^{-\eta \sum_{s=1}^t \ell_t(j)}$ and $W_t = \sum_{j=1}^K W_t(j)$. We have

$$\begin{aligned}
W_t &= \sum_{j=1}^K W_{t-1}(j) e^{-\eta \ell_t(j)} && \leftarrow \quad W_t^{(j)} = W_{t-1}(j) e^{-\eta \ell_t(j)} \\
&= W_{t-1} \sum_{j=1}^K \frac{W_{t-1}(j)}{W_{t-1}} e^{-\eta \ell_t(j)} \\
&= W_{t-1} \sum_{j=1}^K p_t^{(j)} e^{-\eta \ell_t(j)} && \leftarrow \quad p_t(j) = \frac{e^{-\eta \sum_{s=1}^t \ell_t(j)}}{\sum_{k=1}^K e^{-\eta \sum_{s=1}^t \ell_t^{(k)}}} = \frac{W_{t-1}(j)}{W_{t-1}} \\
&\leq W_{t-1} \sum_{j=1}^K p_t(j)\big(1 - \eta \ell_t(j) + \eta^2 \ell_t(j)^2\big) && \leftarrow \quad e^x \leq 1 + x + x^2 \text{ for } x \leq 1 \\
&= W_{t-1}(1 - \eta p_t \cdot \ell_t + \eta^2 p_t \cdot \ell_t^2)\,,
\end{aligned}$$

where we assumed in the inequality $\eta \ell_t(j) \leq 1$ and where we denote $\ell_t = (\ell_t^{(1)}, \dots, \ell_t^{(K)})$, $\ell_t^2 = \big(\ell_t(1)^2, \dots, \ell_t(K)^2\big)$ and $p_t = (p_t(1), \dots, p_t(K))$. Now, using $1 + x \leq e^x$, we get:

$$W_t \leq W_{t-1} \exp\big(-\eta p_t \cdot \ell_t + \eta^2 p_t \cdot \ell_t^2\big)\,.$$

By induction on $t = 1, \dots, T$, this yields using $W_0 = K$

$$W_T \leq K \exp\Big(-\eta \sum_{t=1}^T p_t \cdot \ell_t + \eta^2 \sum_{t=1}^T p_t \cdot \ell_t^2\Big)\,. \tag{2}$$

On the other hand, upper-bounding the maximum with the sum,

$$\exp\Big(-\eta \min_{j \in [K]} \sum_{t=1}^T \ell_t(j)\Big) \leq \sum_{j=1}^K \exp\Big(-\eta \sum_{t=1}^T \ell_t(j)\Big) \leq W_T\,.$$

Combining the above inequality with Inequality (2) and taking the log, we get

$$-\eta \min_{j \in [K]} \sum_{t=1}^T \ell_t(j) \leq -\eta \sum_{t=1}^T p_t \cdot \ell_t + \eta^2 \sum_{t=1}^T p_t \cdot \ell_t^2 + \log K\,. \tag{3}$$

Dividing by η and reorganizing the terms proves the first inequality. Now, we remark that $\mathbb{E}[\ell_t(x_t)] = \mathbb{E}\big[\mathbb{E}\big[\ell_t(x_t)|p_t\big]\big] = \mathbb{E}\big[p_t \cdot \ell_t\big]$ which yields

$$\mathbb{E}\big[R_T\big] = \mathbb{E}\Big[\sum_{t=1}^T p_t \cdot \ell_t - \max_{j \in [K]} \sum_{t=1}^T \ell_t(j)\Big]\,.$$

Substituting into the expectation of Inequality 3 and reorganizing the terms entails

$$\mathbb{E}\big[R_T\big] \leq \frac{\log K}{\eta} + \eta \sum_{t=1}^T \mathbb{E}\big[p_t \cdot \ell_t^2\big]\,. \tag{small losses}$$

as soon as $\eta \ell_t(j) \leq 1$ for all t and j. This last equation is called the "small losses" property since the regret is smaller if $\sum_{t=1}^T \mathbb{E}[p_t \cdot \ell_t^2]$ are small. Optimizing η and upper-bounding $p_t \cdot \ell_t^2 \leq 1$ concludes. □

The constant 2 can be slightly improved to $\sqrt{2}$ (see Cesa-Bianchi and Lugosi [2006]) but otherwise the bound is optimal.

Anytime algorithm (the doubling trick) The previous algorithms EWA depends on a parameter $\eta > 0$ that needs to be optimized according to K and T. For instance, for EWA using the value

$$\eta = \sqrt{\frac{\log K}{KT}} \,.$$

the bound of Theorem 1 is only valid for horizon T. However, the learner might not know the time horizon in advance and one might want an algorithm with guarantees valid simultaneously for all $T \geq 1$. We can avoid the assumption that T is known in advance, at the cost of a constant factor, by using the so-called *doubling trick*. The general idea is the following. Whenever we reach a time step t which is a power of 2, we restart the algorithm (forgetting all the information gained in the past) setting η to $\sqrt{\log K/t}$. Let us denote EWA-doubling this algorithm.

Theorem 2 (Anytime bound on the regret). *For all $T \geq 1$, the pseudo-regret of EWA-doubling is then upper-bounded as:*

$$\mathbb{E}\big[R_T\big] \leq 7\sqrt{T \log K} \,.$$

The same trick can be used to turn EXP3 into anytime algorithms. Actually, we can use the *doubling trick* whenever we have an algorithm with a regret of order $\mathcal{O}(T^\alpha)$ for some $\alpha > 0$ with a known horizon T to turn it into an algorithm with a regret $\mathcal{O}(T^\alpha)$ for all $T \geq 1$.
Another solution is to use time-varying parameters η_t replacing T with the current value of t. The analysis is however less straightforward. Prove as an exercise a regret bound for the choice $\eta_t = \sqrt{\log K/t}$.

Proof. For simplicity we assume $T = 2^{M+1} - 1$. The regret of EWA-doubling is then upper-bounded as:

$$\begin{aligned} R_T &= \mathbb{E}\left[\sum_{t=1}^T \ell_t(x_t) - \min_{i\in[K]} \sum_{t=1}^T \ell_t(i)\right] \\ &\leq \mathbb{E}\left[\sum_{t=1}^T \ell_t(x_t) - \sum_{m=0}^M \min_{i\in[K]} \sum_{t=2^m}^{2^{m+1}-1} \ell_t(i)\right] \\ &= \sum_{m=0}^M \underbrace{\mathbb{E}\left[\sum_{t=2^m}^{2^{m+1}-1} \ell_t(x_t) - \min_{i\in[K]} \sum_{t=2^m}^{2^{m+1}-1} \ell_t(i)\right]}_{R_m} . \end{aligned}$$

Now, we remark that each term R_m corresponds to the expected regret of an instance of EWA over the 2^m rounds $t = 2^m, \dots, 2^{m+1} - 1$ and run with the optimal parameter $\eta = \sqrt{\log K/2^m}$. Therefore, using Theorem 1, we get $R_m \leq 2\sqrt{2^m \log K}$, which yields:

$$R_T \leq \sum_{m=0}^M 2\sqrt{2^m \log K} \leq 2(1+\sqrt{2})\sqrt{2^{M+1}\log K} \leq 7\sqrt{T \log K} \,.$$

□

3 Bandit feedback

In many applications the learner does not observe the losses of all actions but only the one he has chosen. This problem is called multi-armed bandits. Some possible applications are in medicine to find the correct dosage for drugs, in online advertisements or for gps (shortest path).

In this setting, one needs to manage a trade-off between exploration (explore new decision to see how they perform) and exploitation. The historical method was to first explore all decisions during a certain period of time then make a choice regarding the best action and exploit it. But it is possible to explore and exploit simultaneously.

Stochastic Bandits

If the losses ℓ_t are i.i.d. there are three well known methods that address the exploration-exploitation problem:

- UCB (Upper-confidence bound): build confidence intervals for each $\ell_t(k)$ and chooses the action that maximizes the upper-bound.
- ε-Greedy: at each time step, explore uniformly over actions with probability ε or take the action with minimal average loss otherwise.
- Thomson-sampling: choose the action that minimizes the expected loss with respect to some random belief.

For i.i.d. losses, in multi-armed bandits the player can ensure an expected regret of order

$$\mathbb{E}[R_T] \leq \frac{K \log T}{\Delta}$$

where $\Delta = \mathbb{E}\big[\ell_t(k_2) - \ell_t(k_1)\big]$ is the gap between the expected loss of the best action $k_1 = \arg\min_{k\in[K]} \mathbb{E}[\ell_t(k)]$ and the second best action $k_2 = \arg\min_{k\neq k_1} \mathbb{E}[\ell_t(k)]$.

Adversarial bandits

In the following, we will address a problem which seem harder: adversarial losses. Surprisingly, this is still possible to ensure sublinear regret! First, instead of minimizing the *expected regret* $\mathbb{E}[R_T] = \mathbb{E}[\sum_t \ell_t(x_t)] - \mathbb{E}\big[\min_j \sum_t \ell_t(j)\big]$, we consider an easier objective, the *pseudo-regret* defined as

$$\bar{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(x_t)\right] - \min_{j\in[K]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(j)\right]. \qquad \text{(pseudo regret)}$$

Note that $\bar{R}_T \leq \mathbb{E}[R_T]$. If the adversary cannot adapt to the strategy of the learner and the $\ell_t(j)$ are deterministic there is equality.

Ideally, we would like to reuse our algorithm EWA that assigned weights

$$p_t(k) = \frac{e^{-\eta \sum_{s=1}^{t-1} \ell_s(k)}}{\sum_{j=1}^{K} e^{-\eta \sum_{s=1}^{t} \ell_s(j)}}.$$

However this is not possible since we do not observe $\ell_t(k)$ for $k \neq x_t$. The idea is to replace $\ell_t(k)$ with an unbiased estimate that we observe. The first idea would be to use $\ell_t(k)$ if we observe it and 0 otherwise:

$$\widehat{\ell}_t(k) = \ell_t(k) \mathbb{1}_{k=x_t}.$$

However this estimate is biased: the actions that are less likely to be chosen by the algorithm (small weight $p_t(k)$) are more likely to incure 0 loss and we will favorized

$$\mathbb{E}_{x_t\sim p_t}\big[\widehat{\ell}_t(x_t)\big] = p_t(k)\ell_t(k) \ \neq \ell_t(k).$$

We see that we need to correct this phenomenon. Therefore we choose

$$\widehat{\ell}_t(k) = \frac{\ell_t(k)}{p_t(k)} \mathbb{1}_{k=x_t},$$

which leads to the algorithm EXP3 detailed below.

```
EXP3
Parameter: $\eta > 0$
Initialize: $p_1 = \big(\frac{1}{K}, \dots, \frac{1}{K}\big)$
For $t = 1, \dots, T$
    – draw $x_t \sim p_t$; incur loss $\ell_t(x_t)$ and observe $\ell_t(x_t) \in [0,1]$;
    – update for all $k \in \{1, \dots, K\}$
        $p_{t+1}(k) = \frac{e^{-\eta \sum_{s=1}^t \widehat{\ell}_s(k)}}{\sum_{j=1}^K e^{-\eta \sum_{s=1}^t \widehat{\ell}_s(j)}}$,   where $\widehat{\ell}_s(k) = \frac{\ell_s(k)}{p_s(k)} \mathbb{1}_{k=x_s}$
```

Then applying Theorem 1 with the substituted losses $\widehat{\ell}_t$, we get the following theorem.

Theorem 3. *Let $T \geq 1$. The pseudo-regret of EXP3 run with $\eta = \sqrt{\frac{\log K}{KT}}$ is upper-bounded as:*

$$\bar{R}_T \leq 2\sqrt{KT \log K}\,.$$

Proof. Apply EWA to the estimated losses $\widehat{\ell}_t(j)$, we get from the small losses property:

$$\mathbb{E}\bigg[\sum_{t=1}^T \widehat{\ell}_t(x_t) - \min_{j \in [K]} \sum_{t=1}^T \widehat{\ell}_t(j)\bigg] \leq \frac{\log K}{\eta} + \eta \sum_{t=1}^T \mathbb{E}\big[p_t \cdot \widehat{\ell}_t^2\big]\,. \tag{4}$$

Now we compute the expectations. We have $\mathbb{E}\big[\widehat{\ell}_t(j)\big] = \mathbb{E}[\ell_t(j)]$ and

$$\mathbb{E}\big[\widehat{\ell}_t(x_t)\big] = \mathbb{E}\bigg[\sum_{j=1}^K p_t(j)\widehat{\ell}_t(j)\bigg] = \mathbb{E}\bigg[\sum_{j=1}^K p_t(j)\mathbb{E}\big[\widehat{\ell}_t(x_t)|p_t\big]\bigg] = \mathbb{E}\bigg[\sum_{j=1}^K p_t(j)\mathbb{E}\big[\ell_t(j)\big]\bigg] = \mathbb{E}\big[\ell_t(x_t)\big]\,.$$

Therefore, we can lower-bound the left-hand side:

$$\begin{aligned}
\mathbb{E}\bigg[\sum_{t=1}^T \widehat{\ell}_t(x_t) - \min_{j \in [K]} \sum_{t=1}^T \widehat{\ell}_t(j)\bigg] &\geq \max_{j \in [K]} \mathbb{E}\bigg[\sum_{t=1}^T \widehat{\ell}_t(x_t) - \sum_{t=1}^T \widehat{\ell}_t(j)\bigg] \\
&= \max_{j \in [K]} \sum_{t=1}^T \mathbb{E}\big[\widehat{\ell}_t(j) - \widehat{\ell}_t(x_t)\big] = \bar{R}_T\,.
\end{aligned}$$

On the other hand, the expectation of the right-hand side satisfies

$$\begin{aligned}
\mathbb{E}\big[p_t \cdot \widehat{\ell}_t^2\big] = \mathbb{E}\bigg[\sum_{j=1}^K p_t(j)\widehat{\ell}_t(j)^2\bigg] &= \mathbb{E}\bigg[\sum_{j=1}^K p_t(j)\mathbb{E}\big[\widehat{\ell}_t(j)^2|p_t\big]\bigg] \\
&= \mathbb{E}\bigg[\sum_{j=1}^K \sum_{k=1}^K p_t(j)p_t(k)\Big(\frac{\ell_t(j)}{p_t(j)}\mathbb{1}\{j=k\}\Big)^2\bigg] \\
&= \mathbb{E}\bigg[\sum_{j=1}^K \sum_{k=1}^K p_t(k)\frac{\ell_t(j)^2}{p_t(j)}\mathbb{1}\{j=k\}\bigg] \\
&= \mathbb{E}\bigg[\sum_{j=1}^K \ell_t(j)^2\bigg] \leq K\,.
\end{aligned}$$

Substituting into Inequality (4) yields

$$\bar{R}_T \leq \frac{\log K}{\eta} + \eta K T\,.$$

and optimizing $\eta = \sqrt{dT/(\log K)}$ concludes. □

A bound on the true regret R_T instead of the pseudo-regret $\bar{R}_T$ can be obtained by another algorithm EXP3.P (see Cesa-Bianchi and Lugosi [2006]). The issue of EXP3 is that it is very unstable because the estimated losses can be very large when $p_t(k)$ is small. EXP3.P requires an exploration parameter γ to ensure $p_t(k) > \gamma$ for all actions.

The rate $O(\sqrt{KT})$ is optimal for bandit feedback. We only loose a factor $\sqrt{K}$ with respect to full information! Intuitively, this is because we observe K times less information and thus T is changed into KT.

References

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games.* Cambridge university press, 2006.

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.