Adversarial Bandits

Pierre Gaillard

November 2019

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 in Figure 1.

At each time step $t = 1, \dots, T$
- the player chooses an action $x_t \in \mathcal{X}$ (compact decision set);
- 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 $x \in \mathcal{X}$ $\quad\to\quad$ full-information feedback
  - the loss of the chosen action only: $\ell_t(x_t)$ $\quad\to\quad$ bandit feedback.

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

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

Figure 1: Setting of an online learning problem

Example 1.1 (Multi-armed bandit)**.** In K-armed bandit, the decision set are K actions (or arms) $\mathcal{X} = \{1, \dots, K\}$ and the player only observes the performance of the chosen action (bandit feedback). In this problem, there is an exploration-exploitation trade-off: the player wants to select the best arm as often as possible but he also needs to explore all arms to estimate their performance.

This problem takes his name from slot machines (also known as one-armed bandits because they were originally operated by one lever on the side of the machine) in which some player explores several slot machines and tries to maximize his cumulative gain (or more likely minimize his loss!).

Originally, multi-armed bandit setting was introduced by Thomson in 1933 and motivated by clinical trials. For the tth patient in some clinical study, one needs to choose the treatment to assign to this patient and observe the response. The goal is to maximize the number of patients healed during the study.

Nowadays, multi-armed bandit is motivated by many applications coming from internet (recommender systems, online advertisements,...).

Example 1.2 (Prediction with expert advice)**.** In prediction with expert advice, there is some sequence of observations $y_1, \dots, y_T \in [0,1]$ to be predicted step by step with the help of expert forecasts. The setting can be formalized as follows: at each time step $t \geq 1$

- the environment reveals experts forecasts $f_t(k)$ for $k = 1, \dots, K$
- the player chooses a weight vector $p_t \in \Delta_K \stackrel{\text{def}}{=} \{p \in [0,1]^K : \sum_{k=1}^K p_k = 1\}$
- the player forecasts $\widehat{y}_t = \sum_{k=1}^K p_t(k) f_t(k)$
- the environment reveals $y_t \in [0,1]$ and the player suffers loss $\ell_t(p_t) = \ell(\widehat{y}_t, y_t)$ where $\ell : [0,1]^2 \to [0,1]$ is a loss function.

Considering $\mathcal{X} = \Delta_K$ and $x_t = p_t$, this setting can be recovered by the online learning setting of Figure 1 with the small difference that the expert advice $f_t(k)$ are often revealed before the learner makes his decision p_t.

Player's performance is then measured via a loss function $\ell_t(x_t) = \ell(\widehat{y}_t, y_t)$ which measures the distance between the prediction $\widehat{y}_t$ and the output y_t. Typical loss functions are the square loss $\ell(\widehat{y}_t, y_t) = (\widehat{y}_t - y_t)^2$, the absolute loss $\ell(\widehat{y}_t, y_t) = |\widehat{y}_t - y_t|$ or the absolute percentage of error $\ell(\widehat{y}_t, y_t) = |\widehat{y}_t - y_t|/|y_t|$. All these loss functions are convex, which will play an important role in the analysis.

How to measure the performance: the regret 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*

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

We have some bias-variance decomposition:

$$\sum_{t=1}^T \ell_t(x_t) \;=\; \underbrace{\inf_{x \in \mathcal{X}} \sum_{t=1}^T \ell_t(x)}_{\text{Approximation error = how good the possible actions are.}} \;+\; \underbrace{R_T}_{\text{Sequential estimation error of the best action}}$$

We will focus on the regret in these lectures. 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.

Remark. In this lesson, we will not make any random assumption on the process generating the losses ℓ_t. The latter are deterministic and may be chosen by some adversary.

Remark. Note that the loss functions ℓ_t depend on the round t. This may be caused by many phenomena. We provide here some possible reasons. This may be because

- of some observation to be predicted if $\ell_t(x) = \ell(x, y_t)$. For instance, if the goal is to predict the evolution of the temperature $y_1, \dots, y_T$, the latter changes over time and a prediction x is evaluated with $\ell_t(x) = (x - y_t)^2$.

– the environment is stochastic and the variation over time t models some noise effect.
– of a changing environment. For instance, if the player is playing a game against some adversary that evolves and adapts to its strategy. A typical example is the case of spam detections. If the player tries to detect spams, while some spammers (the environment) try at the same time to fool the player with new spam strategies.

2 Full information feedback

We will start with the simple case of the simplex as decision set $\mathcal{X} = \Delta_K = \{p \in [0,1]^K : \sum_{k=1}^K p_k = 1\}$ (in this part we will call p_t the decision of the learner instead of x_t) with the linear loss function

$$\forall p \in \mathcal{X}, \qquad \ell_t(p) = \sum_{k=1}^K p(k)\ell_t(k)$$

where $\ell_t = (\ell_t(1), \dots, \ell_t(K)) \in [0,1]^K$ is a loss vector chosen by the environment at round t. Note that for simplicity there is an abuse of notation $\ell_t(k) = \ell_t(\delta_K)$. Remark that in this linear setting

$$\min_{p\in\mathcal{X}} \sum_{t=1}^T \ell_t(p) = \min_{p\in\mathcal{X}} \sum_{t=1}^T p \cdot \ell_t = \min_{p\in\mathcal{X}} p \cdot \Big(\sum_{t=1}^T \ell_t\Big) = \min_{1\le k\le K} \sum_{t=1}^T \ell_t(k)\,.$$

The best fixed weight vector p is a Dirac mass on the best fixed decision $k \in \{1, \dots, K\}$.

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 it would not work (see exercises). 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 $t = 1, \dots, T$
– select p_t; incur loss $\ell_t(p_t)$ and observe $\ell_t \in [0,1]^K$;
– update for all $k \in \{1, \dots, K\}$

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

Exercise 2.1. Consider the strategy, called "Follow The Leader" (FTL) that puts all the mass on the best action so far:

$$p_t \in \arg\min_{p\in\mathcal{X}} \sum_{s=1}^{t-1} \ell_s(p)\,. \tag{FTL}$$

1. Show that $p_t(k) > 0$ implies that $k \in \arg\min_j \sum_{s=1}^{t-1} \ell_s(j)$
2. Show that the regret of FTL might be linear: i.e., there exists a sequence $\ell_1, \dots, \ell_T \in [0,1]^K$ and $c > 0$ such that $R_T \ge cT$.

The following theorem proves that EWA, which is a smoothed version of FTL, achieves sublinear regret.

Theorem 1. *Let $T \ge 1$. For all sequences of losses $\ell_1, \dots, \ell_T \in [0,1]^K$, EWA achieves the bound*

$$R_T \overset{def}{=} \sum_{t=1}^T p_t \cdot \ell_t - \min_{1\le j\le K} \sum_{t=1}^T \ell_t(j) \le \eta \sum_{t=1}^T \sum_{k=1}^K p_t(k)\ell_t(k)^2 + \frac{\log K}{\eta}\,. \tag{1}$$

Therefore, for the choice $\eta = \sqrt{\frac{\log K}{T}}$*, EWA satisfies the regret bound* $R_T \leq 2\sqrt{T \log K}$ *.*

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

Exercise 2.2. Generalize the above theorem when the losses $\ell_1, \dots, \ell_T \in [-B, B]$ for some $B > 0$.

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:

$$R_T \overset{\text{def}}{=} \sum_{t=1}^T p_t \cdot \ell_t - \min_{1 \leq j \leq K} \sum_{t=1}^T \ell_t(j) \leq \eta \sum_{t=1}^T p_t \cdot \ell_t^2 + \frac{\log K}{\eta}$$

Optimizing η and upper-bounding $p_t \cdot \ell_t^2 \leq 1$ concludes the second inequality. □

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:*

$$R_T \leq 7\sqrt{T \log K}\,.$$

The same trick can be used to turn most online algorithms into anytime algorithms (even in more general settings: bandits, general loss,...). 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.

Exercise 2.3. Prove a regret bound for the time-varying choice $\eta_t = \sqrt{\log K/t}$ in EWA.

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

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

Now, we remark that each term R_m corresponds to the expected regret of an instance of EWA over the 2^m rounds $t = 2^m, \ldots, 2^{m+1} - 1$ 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:

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

□

The first inequality in Theorem 1 is often called improvement for small losses since the regret is smaller if $\sum_{t=1}^{T} p_t \cdot \ell_t^2 \leq \sum_{t=1}^{T} p_t \cdot \ell_t = \widehat{L}_t$ are small.

3 Adversarial bandits

Now, we will see adversarial bandits: that is bandit feedback (only $\ell_t(x_t)$ is observed) with an adversarial sequence of loss function ℓ_t (i.e., no stochastic assumptions). Note that we turn back to losses instead of rewards but we will come back to rewards whenever it makes the proof easier. Remember that the lower-bound on the regret in the worst-case is of order $O(\sqrt{TK})$.

We consider Setting 1 with bandit feedback, finite decision space $\mathcal{X} \overset{\text{def}}{=} \{1, \dots, K\}$ and adversarial losses. We do not assume the loss functions ℓ_t to be linear nor convex (the decision space is not). similarly to Random EWA the chosen action $x_t \in \mathcal{X}$ is sampled randomly from a distribution p_t chosen at round t by the player. We will provide an algorithm called Exp3 inspired by EWA.

Let us denote the regret with respect to action $k \in \mathcal{X}$ by

$$R_T(k) \overset{\text{def}}{=} \sum_{t=1}^{T} \ell_t(x_t) - \sum_{t=1}^{T} \ell_t(k)\,.$$

Instead of minimizing the *expected regret* $\mathbb{E}[R_T] = \mathbb{E}[\max_k R_T(k)]$, we will consider an easier objective, the *pseudo-regret* defined as

$$\bar{R}_T \overset{\text{def}}{=} \max_{k\in\mathcal{X}} \mathbb{E}\big[R_T(k)\big] = \max_{k\in\mathcal{X}} \mathbb{E}\bigg[\sum_{t=1}^{T} \ell_t(x_t) - \sum_{t=1}^{T} \ell_t(k)\bigg]\,. \qquad \text{(pseudo regret)}$$

It is worth pointing out that the expectations are taken with respect to the randomness of the algorithm: the decisions x_t are random. We can distinguish two types of adversaries:

- *oblivious adversary*: all the loss functions $\ell_1, \dots, \ell_T$ are chosen in advance before the game starts and do not depend on the past player decisions $x_1, \dots, x_T$. In this case, the losses $\ell_t(k)$ are determinist and there is thus equality: $\bar{R}_T = \mathbb{E}[R_T]$.
- *adaptive adversary*: the loss function ℓ_t at round $t \geq 1$ may depend on past information $\sigma(x_1, \dots, x_{t-1})$. It is thus random. By Jensen's inequality $\max_{k\in\mathcal{X}} \mathbb{E}\big[R_T(k)\big] \leq \mathbb{E}\big[\max_{k\in\mathcal{X}} R_T(k)\big]$ and thus $\bar{R}_T \leq \mathbb{E}[R_T]$.

The EXP3 algorithm Ideally, we would like to reuse our algorithm EWA that assigned weights

$$\forall k \in \mathcal{X}, \qquad p_t(k) = \frac{e^{-\eta \sum_{s=1}^{t-1} \ell_s(k)}}{\sum_{j=1}^{K} e^{-\eta \sum_{s=1}^{t} \ell_s(j)}}\,. \qquad \text{(EWA)}$$

Unfortunately this is not possible since the player does not observe $\ell_t(k)$ for $k \neq x_t$. The high-level idea of Exp3 is to replace $\ell_t(k)$ with an unbiased estimate that is observed by the player. A first idea would be to use $\ell_t(k)$ if we observe it and 0 otherwise:

$$\widehat{\ell}_t(k) = \begin{cases} \ell_t(k) & \text{if } k = x_t \quad \leftarrow \text{i.e., decision } k \text{ is observed} \\ 0 & \text{otherwise} \end{cases}\,.$$

However, this estimate is biased:

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

In other words, the actions that are less likely to be chosen by the algorithm (small weight $p_t(k)$) are more likely to be unobserved and incur 0 loss. We need to correct this phenomenon. Therefore we choose

$$\widehat{\ell}_t(k) = \frac{\ell_t(k)}{p_t(k)} \mathbb{1}_{k=x_t}\,, \qquad (4)$$

which leads to the algorithm EXP3 detailed below.

EXP3

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 $\ell_t(x_t) \in [0,1]$;
- update for all $k \in \{1, \dots, K\}$

$$p_{t+1}(k) = \frac{e^{-\eta \sum_{s=1}^{t} \widehat{\ell}_s(k)}}{\sum_{j=1}^{K} e^{-\eta \sum_{s=1}^{t} \widehat{\ell}_s(j)}}, \quad \text{where } \widehat{\ell}_s(k) = \frac{\ell_s(k)}{p_s(k)} \mathbb{1}_{k=x_s}$$

Then applying the Inequality (1) for EWA 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 $\widehat{\ell}_t(j)$ that are completely observed (nonnegative but not bounded), we get from Inequality (1) and taking the expectation:

$$\mathbb{E}\left[\sum_{t=1}^{T} p_t \cdot \widehat{\ell}_t - \min_{j \in \mathcal{X}} \sum_{t=1}^{T} \widehat{\ell}_t(j)\right] \leq \frac{\log K}{\eta} + \eta \sum_{t=1}^{T} \mathbb{E}\Big[p_t \cdot \widehat{\ell}_t^2\Big]\,. \tag{5}$$

Now we compute the expectations. Denote by $\mathcal{F}_{t-1} \stackrel{\text{def}}{=} \sigma(p_1, \ell_1, x_1, \dots, x_{t-1}, p_t, \ell_t)$ the past information available at round t for the adversary (which cannot use the randomness of x_t but can use p_t). Note that ℓ_t and p_t are $\mathcal{F}_{t-1}$-measurable by assumption. We have

$$\forall j \in \mathcal{X} \qquad \mathbb{E}\Big[\widehat{\ell}_t(j)\Big|\mathcal{F}_{t-1}\Big] = \mathbb{E}\Big[\frac{\ell_t(j)}{p_t(j)}\mathbb{1}_{j=x_t}\Big|\mathcal{F}_{t-1}\Big] = \sum_{k=1}^{K} p_t(k)\frac{\ell_s(j)}{p_t(j)}\mathbb{1}_{j=k} = \ell_t(j)$$

thus the estimated losses are unbiased $\mathbb{E}\big[\widehat{\ell}_t(j)\big] = \mathbb{E}\big[\ell_t(j)\big]$ and

$$\begin{aligned}\mathbb{E}\Big[p_t \cdot \widehat{\ell}_t\Big] = \mathbb{E}\Big[\sum_{j=1}^{K} p_t(j)\widehat{\ell}_t(j)\Big] &= \mathbb{E}\Big[\sum_{j=1}^{K} p_t(j)\mathbb{E}\Big[\widehat{\ell}_t(j)\Big|\mathcal{F}_{t-1}\Big]\Big] \\ &= \mathbb{E}\Big[\sum_{j=1}^{K} p_t(j)\ell_t(j)\Big] = \mathbb{E}\Big[\mathbb{E}\big[\ell_t(x_t)\big|\mathcal{F}_{t-1}\big]\Big] = \mathbb{E}\big[\ell_t(x_t)\big]\,.\end{aligned}$$

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

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

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

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

Substituting into Inequality (5) yields

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

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

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.