Adversarial bandits

Pierre Gaillard

Reminder from last lectures

We recall the setting of online convex optimization in Figure 1.

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

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

$$R_T \overset{\text{def}}{=} \sum_{t=1}^{T} \ell_t(\theta_t) - \min_{\theta\in\Theta} \sum_{t=1}^{T} \ell_t(\theta)\,,$$

for any possible loss sequence $\ell_1, \dots, \ell_t$ chosen by the environment.

Setting 1: Setting of an online learning problem

In previous lectures, we considered the full-information feedback and the bandit feedback with stochastic loss functions. In *full information with finite decision set* $\Theta = [K] \overset{\text{def}}{=} \{1, \dots, K\}$, we saw the Random Exponentially Weighted Average (EWA) forecaster. It is defined as

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

and draws $\theta_t = k$ with probability $p_t(k)$. If $-\eta\ell_t(j) \le 1$ (see the proof of EWA in first lecture), it satisfies the upper-bound:

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

Since the decision θ_t is random, we assume that ℓ_t cannot depend on θ_t but may depend on past information $\sigma(p_1, \ell_1, x_1, \dots, x_{t-1}, p_t)$. The above bound can be converted into a bound on the expected regret for well-calibrated learning rate η

$$\mathbb{E}\big[R_T\big] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\theta_t) - \min_{k\in[K]} \sum_{t=1}^{T} \ell_t(k)\right] \le 2\sqrt{T \log K}\,.$$

In this lecture, we will see adversarial bandits: that is bandit feedback (only $\ell_t(\theta_t)$ 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 $O(\sqrt{TK})$.

1 The exponentially weighted average algorithm for bandits

We consider Setting 1 with bandit feedback, finite decision space $\Theta = [K] \stackrel{\text{def}}{=} \{1, \dots, K\}$ and adversarial losses. To emphasize that the action is in $[K]$, we denote by k_t the action chosen by the player (instead of θ_t). 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 $k_t \in [K]$ 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.

1.1 Pseudo-regret bound

Let us denote the regret with respect to action $k \in [K]$ by

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

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

$$\bar{R}_T \stackrel{\text{def}}{=} \max_{k\in[K]} \mathbb{E}\big[R_T(k)\big] = \max_{k\in[K]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(k_t) - \sum_{t=1}^{T} \ell_t(k)\right]. \qquad \text{(pseudo regret)}$$

It is worth pointing out that the expectations are taken with respect to the randomness of the algorithm: the decisions k_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 $k_1, \dots, k_T$. 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 $\sigma(k_1, \dots, k_{t-1})$. It is thus random. By Jensen's inequality $\max_{k\in[K]} \mathbb{E}\big[R_T(k)\big] \leq \mathbb{E}\big[\max_{k\in[K]} R_T(k)\big]$ 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

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

Unfortunately this is not possible since the player does not observe $\ell_t(k)$ for $k \neq k_t$. 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:

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

However, this estimate is biased:

$$\mathbb{E}_{k_t \sim p_t}\big[g_t(k_t)\big] = p_t(k)\ell_t(k) \;\neq \ell_t(k)\,.$$

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

$$g_t(k) = \frac{\ell_t(k)}{p_t(k)} \mathbb{1}_{\{k=k_t\}}\,, \qquad (1)$$

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 $k_t \sim p_t$; incur loss $\ell_t(k_t)$ and observe $\ell_t(k_t) \in [0,1]$;
- update for all $k \in \{1, \dots, K\}$

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

Then applying the Inequality $(*)$ for EWA with the substituted losses g_t, we get the following theorem.

Theorem 1. *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 $g_t(j)$ that are completely observed (nonnegative but not bounded), we get from Inequality $(*)$ and taking the expectation:

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

Now we compute the expectations. Denote by $\mathcal{F}_{t-1} \stackrel{\text{def}}{=} \sigma(p_1, \ell_1, k_1, \dots, k_{t-1}, p_t, \ell_t)$ the past information available at round t for the adversary (which cannot use the randomness of k_t but can use p_t). Note that ℓ_t and p_t are $\mathcal{F}_{t-1}$-measurable by assumption. We have

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

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

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

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

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

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

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

Substituting into Inequality (2) yields

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

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

The issue with the above regret bound is that it bounds the pseudo-regret and not the expected regret. This is because we have

$$\mathbb{E}\Big[\min_j \sum_{t=1}^T g_t(j)\Big] \le \min_j \mathbb{E}\Big[\sum_{t=1}^T g_t(j)\Big] = \min_{j\in[K]} \mathbb{E}\Big[\sum_{t=1}^T \ell_t(j)\Big]$$

but not

$$\mathbb{E}\Big[\min_j \sum_{t=1}^T g_t(j)\Big] \not\le \mathbb{E}\Big[\min_j \sum_{t=1}^T \ell_t(j)\Big]\,. \tag{3}$$

Hence, controlling the cumulative loss agains the best estimated action only controls the pseudo regret and not the true regret.

1.2 High probability bound on the regret

Gains versus losses In this part, we will switch the analysis from losses $\ell_t(k)$ to gains $g_t(k) = 1 - \ell_t(k) \in [0,1]$ because the core idea of the next algorithm is easier to see with gains. Remark that the loss and gain versions are symmetric via the transformation $g_t(k) = 1 - \ell_t(k)$. The regret in terms of gains is defined as

$$R_T \overset{\text{def}}{=} \max_{k\in[K]} \sum_{t=1}^T g_t(k) - \sum_{t=1}^T g_t(k_t)\,.$$

Using EWA with full information from $(*)$, if $\eta g_t(k) \le 1$, we also have for gains the inequality

$$\max_{1\le j\le K} \sum_{t=1}^T g_t(j) - \sum_{t=1}^T p_t \cdot g_t \le \eta \sum_{t=1}^T p_t \cdot g_t^2 + \frac{\log K}{\eta}\,, \quad \text{where} \quad p_t(k) = \frac{e^{\eta \sum_{s=1}^{t-1} g_s(k)}}{\sum_{j=1}^K e^{\eta \sum_{s=1}^{t-1} g_s(j)}}\,. \tag{4}$$

High-level idea of EXP3.P The high-level idea of the next algorithm is to ensure that the estimators $\widehat{g}_t(k)$ of the gains satisfy

$$\mathbb{E}\left[\max_j \sum_{t=1}^T \widehat{g}_t(j)\right] \ge \mathbb{E}\left[\max_j \sum_{t=1}^T g_t(j)\right] \tag{5}$$

so that controlling the performance with respect to the estimated gains (left-hand side) also controls the performance with respect to the true gains (right-hand side). This was not the case of the estimators used for EXP3 (see (3)). To ensure (5), we add a bias term β to the estimators $\widehat{g}_t(k)$ as follows:

$$\widehat{g}_t(k) \stackrel{\text{def}}{=} \frac{g_t(k)\mathbb{1}_{\{k=k_t\}} + \beta}{p_t(k)} \tag{6}$$

In contrary to (1), the estimator is indeed biased

$$\mathbb{E}\big[\widehat{g}_t(k)\big|\mathcal{F}_{t-1}\big] = g_t(k) + \frac{\beta}{p_t(k)}\,, \tag{7}$$

where we recall that $\mathcal{F}_{t-1} \stackrel{\text{def}}{=} \sigma\big(p_1, k_1, g_1, \ldots, k_{t-1}, p_t, g_t\big)$ contains the information up to time t available to the environment. We have the following Lemma:

Lemma 1. *For any $\delta > 0$, with probability $1-\delta$ and $\beta \in (0,1)$,*

$$\sum_{t=1}^{T} \widehat{g}_t(j) \geq \sum_{t=1}^{T} g_t(j) - \frac{\log(1/\delta)}{\beta}\,.$$

Proof. Let $\beta \in (0,1)$, from Markov's inequality, we have

$$\begin{aligned}
\mathbb{P}\bigg(\sum_{t=1}^{T} \widehat{g}_t(j) \geq \sum_{t=1}^{T} g_t(j) - \frac{\log(1/\delta)}{\beta}\bigg) &= \mathbb{P}\bigg(\exp\Big(\beta\sum_{t=1}^{T}\big(g_t(j) - \widehat{g}_t(j)\big)\Big) \geq \delta^{-1}\bigg) \\
&\leq \delta\mathbb{E}\bigg[\exp\Big(\beta\sum_{t=1}^{T}\big(g_t(j) - \widehat{g}_t(j)\big)\Big)\bigg]\,.
\end{aligned}$$

It only remains to upper-bound the expectation in the right-hand side by 1, which we do now. Since $\beta \in (0,1)$ and $\widehat{g}_t(j) \geq \beta/p_t(j)$, we have $\beta(g_t(j) - \widehat{g}_t(j) + \beta/p_t(j)) \leq 1$. Therefore, we can use the inequality $e^x \leq 1 + x + x^2$ for $x \leq 1$, which entails

$$\begin{aligned}
\mathbb{E}\bigg[\exp\Big(\beta\big(g_t(j) - \widehat{g}_t(j)\big)\Big)\bigg|\mathcal{F}_{t-1}\bigg] &= \mathbb{E}\bigg[\exp\Big(\beta\Big(g_t(j) - \widehat{g}_t(j) + \frac{\beta}{p_t(j)}\Big)\Big)\bigg|\mathcal{F}_{t-1}\bigg]\exp\Big(-\frac{\beta^2}{p_t(j)}\Big) \\
&\leq \mathbb{E}\bigg[\bigg(1 + \cancel{\beta\Big(g_t(j) - \widehat{g}_t(j) + \frac{\beta}{p_t(j)}\Big)} + \beta^2\Big(g_t(j) - \widehat{g}_t(j) + \frac{\beta}{p_t(j)}\Big)^2\bigg)\bigg|\mathcal{F}_{t-1}\bigg]e^{-\frac{\beta^2}{p_t(j)}} \\
&\stackrel{(7)}{=} \bigg(1 + \beta^2\mathbb{E}\bigg[\Big(g_t(j) - \widehat{g}_t(j) + \frac{\beta}{p_t(j)}\Big)^2\bigg|\mathcal{F}_{t-1}\bigg]\bigg)e^{-\frac{\beta^2}{p_t(j)}}
\end{aligned}$$

where the last equality is by (7) and because $p_t(j)$ is $\mathcal{F}_{t-1}$-measurable. Now,

$$\begin{aligned}
\mathbb{E}\bigg[\Big(g_t(j) - \widehat{g}_t(j) + \frac{\beta}{p_t(j)}\Big)^2\bigg|\mathcal{F}_{t-1}\bigg] &= \mathrm{Var}\Big(\widehat{g}_t(j)\Big|\mathcal{F}_{t-1}\Big) = \mathrm{Var}\Big(\frac{g_t(j)\mathbb{1}_{\{j=k_t\}}}{p_t(j)}\Big|\mathcal{F}_{t-1}\Big) \\
&\leq \mathbb{E}\left[\left(\frac{g_t(j)\mathbb{1}_{\{j=k_t\}}}{p_t(j)}\right)^2\bigg|\mathcal{F}_{t-1}\right] \leq \mathbb{E}\left[\frac{\mathbb{1}_{\{j=k_t\}}}{p_t(j)^2}\bigg|\mathcal{F}_{t-1}\right] = \sum_{k=1}^{K}\frac{p_t(k)\mathbb{1}_{\{j=k\}}}{p_t(j)^2} = \frac{1}{p_t(j)}\,.
\end{aligned}$$

Substituting into the previous inequality and using $1 + x \leq e^x$, it yields

$$\mathbb{E}\bigg[\exp\Big(\beta\big(g_t(j) - \widehat{g}_t(j)\big)\Big)\bigg|\mathcal{F}_{t-1}\bigg] \leq \Big(1 + \frac{\beta^2}{p_t(j)}\Big)e^{-\beta^2/p_t(j)} \leq 1\,.$$

The proof is concluded by induction

$$\mathbb{E}\Big[\exp\Big(\beta\sum_{t=1}^{T}\big(g_t(j)-\widehat{g}_t(j)\big)\Big)\Big] = \mathbb{E}\Big[\underbrace{\mathbb{E}\Big[\exp\Big(\beta\big(g_T(j)-\widehat{g}_T(j)\big)\Big)\Big|\mathcal{F}_{T-1}\Big]}_{\leq 1}\exp\Big(\beta\sum_{t=1}^{T-1}\big(g_t(j)-\widehat{g}_t(j)\big)\Big)\Big]$$

$$\leq \mathbb{E}\Big[\exp\Big(\beta\sum_{t=1}^{T-1}\big(g_t(j)-\widehat{g}_t(j)\big)\Big)\Big] \leq \ldots \leq 1\,.$$

□

The issue with the estimators $\widehat{g}_t(j) \in (0,+\infty)$ defined in Equation (6) is that they might be unbounded if the weights $p_t(j)$ are close to zero. The condition $\eta\widehat{g}_t(j) \leq 1$ which appeared in the proof of EWA cannot hold for any $\eta > 0$. Remark that this was not a problem for EXP3 with the preceding choice (1) because $-\eta g_t(j) \leq 1$ (see the proof of EWA for details).

The next algorithm called EXP3.P, is close to EXP3 but ensures the weights do not vanish to zero by adding an exploration parameter $\gamma > 0$.

EXP3.P

Parameters: $\eta > 0, \beta \in (0,1), \gamma \in (0,1)$
Initialize: $p_1 = \big(\frac{1}{K},\ldots,\frac{1}{K}\big)$
For $t = 1,\ldots,T$
- draw $k_t \sim p_t$; receive gain $g_t(k_t) = 1-\ell_t(k_t)$ and observe $g_t(k_t) \in [0,1]$;
- update for all $k \in \{1,\ldots,K\}$

$$p_{t+1}(k) = (1-\gamma)\frac{e^{\eta\sum_{s=1}^{t}\widehat{g}_s(k)}}{\sum_{j=1}^{K}e^{\eta\sum_{s=1}^{t}\widehat{g}_s(j)}} + \frac{\gamma}{K}\,,$$

where $\widehat{g}_s(k) = \frac{g_s(k)\mathbb{1}_{\{k=k_s\}}+\beta}{p_s(k)}$.

The weights $p_t(k)$ of EXP3.P are necessary larger than γ/K and thus $|\eta g_t(j)| \leq 1$ as soon as $\eta(1+\beta)K/\gamma \leq 1$. We get the following high-probability bound on the regret.

Theorem 2. *For well-chosen parameters $\gamma \in (0,1)$, $\beta \in (0,1)$ and $\eta > 0$ satisfying $\eta(1+\beta)K/\gamma \leq 1$, for any $\delta > 0$, the EXP3.P algorithm achieves*

$$R_T \leq 6\sqrt{TK\log K} + \sqrt{\frac{TK}{\log K}}\log(1/\delta)\,.$$

with probability at least $1-\delta$.

Remark that the above bound leads to a bound on the expected regret, with the choice $\delta = 1/T$ it yields

$$\mathbb{E}\big[R_T\big] \leq 6\sqrt{TK\log K} + \sqrt{\frac{TK}{\log K}}\log(T) + 1$$

The logarithmic dependency on T can even be removed using $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq \varepsilon)d\varepsilon$.

Proof of Theorem 2. Defining the weights that would assign EXP3,

$$q_t(j) \overset{\text{def}}{=} \frac{e^{\eta\sum_{s=1}^{t-1}\widehat{g}_s(j)}}{\sum_{k=1}^{K}e^{\eta\sum_{s=1}^{t-1}\widehat{g}_s(k)}}\,,$$

we get from Inequality (4) applied with $\widehat{g}_t(j)$,

$$\max_{j\in[K]} \sum_{t=1}^{T} \widehat{g}_t(j) \le \sum_{t=1}^{T} q_t \cdot \widehat{g}_t + \eta \sum_{t=1}^{T} q_t \cdot \widehat{g}_t^2 + \frac{\log K}{\eta}\,.$$

where we used $\eta\widehat{g}_t(j) \le 1$ because $\eta(1+\beta)K/\gamma \le 1$. Now, we use that $p_t \overset{\text{def}}{=} (1-\gamma)q_t + \gamma/K$, which entails $q_t = (p_t - \gamma/K)/(1-\gamma) \le p_t/(1-\gamma)$. Substituting into the above inequality

$$(1-\gamma)\max_{j\in[K]} \sum_{t=1}^{T} \widehat{g}_t(j) \le \sum_{t=1}^{T} p_t \cdot \widehat{g}_t + \eta \sum_{t=1}^{T} p_t \cdot \widehat{g}_t^2 + \frac{\log K}{\eta}\,. \tag{8}$$

But by definition of $\widehat{g}_t$,

$$p_t \cdot \widehat{g}_t = \sum_{j=1}^{K} p_t(j)\widehat{g}_t(j) = \sum_{j=1}^{K} \big(g_t(j)\mathbb{1}_{\{j=k_t\}} + \beta\big) = g_t(k_t) + K\beta\,.$$

and since $p_t(j)\widehat{g}_t(j) \le (1+\beta)$,

$$\sum_{t=1}^{T} p_t \cdot \widehat{g}_t^2 \le (1+\beta)\sum_{j=1}^{K}\sum_{t=1}^{T} \widehat{g}_t(j) \le K(1+\beta)\max_{j\in[K]}\sum_{t=1}^{T}\widehat{g}_t(j) \le \frac{\gamma}{\eta}\max_{j\in[K]}\sum_{t=1}^{T}\widehat{g}_t(j)\,.$$

Therefore, substituting into Inequality (8) gives

$$(1-\gamma)\max_{j\in[K]} \sum_{t=1}^{T} \widehat{g}_t(j) \le \sum_{t=1}^{T} g_t(k_t) + K\beta T + \gamma \max_{j\in[K]}\sum_{t=1}^{T}\widehat{g}_t(j) + \frac{\log K}{\eta}\,,$$

where we used $(1+\beta)K \le \gamma/\eta$. Reorganizing, we get

$$(1-2\gamma)\max_{j\in[K]} \sum_{t=1}^{T} \widehat{g}_t(j) \le \sum_{t=1}^{T} g_t(k_t) + K\beta T + \frac{\log K}{\eta}\,.$$

Using Lemma 1 together with a union bound (to have it for all $j \in [K]$), we have with probability $1-\delta$

$$(1-2\gamma)\bigg(\max_{j\in[K]}\sum_{t=1}^{T} g_t(j) - \frac{\log(K/\delta)}{\beta}\bigg) \le \sum_{t=1}^{T} g_t(k_t) + K\beta T + \frac{\log K}{\eta}\,,$$

and thus reorganizing and choosing $\gamma \overset{\text{def}}{=} 2\eta K \ge \eta(1+\beta)K$,

$$\max_{j\in[K]}\sum_{t=1}^{T} g_t(j) - \sum_{t=1}^{T} g_t(k_t) \le K\beta T + \frac{\log K}{\eta} + \frac{\log(K/\delta)}{\beta} + 4\eta K T\,.$$

The proof is concluded by optimizing $\eta \overset{\text{def}}{=} (1/2)\sqrt{(\log K)/KT}$ and $\beta \overset{\text{def}}{=} \sqrt{(\log K)/(KT)}$. □

2 Adversarial bandits with experts

We turn back to the loss version of the game. We now consider prediction with expert advice in the bandit framework. The setting is the same as the one described in Figure 1, but at the beginning of each round $t \ge 1$, some experts $i = 1, \dots, N$ propose recommendations $h_t(i) \in [K]$. These recommendations may be random and may depend on past actions k_s, $s \le t-1$ and past observations $\ell_s(k_s)$. The loss of each expert

is given by the loss of the chosen decision $\ell_t\big(h_t(i)\big)$ but only $\ell_t(k_t)$ is observed by the learner. The goal of the learner is then to be competitive with the best expert on a long run. To do so, it minimizes the pseudo-regret

$$R_T^{\exp} \stackrel{\text{def}}{=} \max_{i=1,\dots,N} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(k_t) - \sum_{t=1}^{T} \ell_t\big(h_t(i)\big)\right]$$

with respect to the experts. In order to bound the pseudo-regret, one could consider experts as the set of arms and use EXP3. This would give a bound of order $\sqrt{TN \log N}$. However it does not take into account the information on the reward of all experts that choose the same action $h_t(i) = k_t$.

EXP4

Parameter: $\eta > 0$
Initialize: $q_1 = \big(\frac{1}{N}, \dots, \frac{1}{N}\big)$.
For each round $t = 1, \dots, n$

1. Get expert advice $h_t(1), \dots, h_t(N) \in [K]$
2. Draw an expert i_t with probability distribution $q_t \in \Delta_N$
3. Choose decision $k_t = h_t(i_t)$
4. Compute the estimated loss for each decision

$$g_t(k) = \frac{\ell_t(k)}{p_t(k)} \mathbb{1}_{\{k=k_t\}}\,,$$

   where $p_t \stackrel{\text{def}}{=} \sum_{i=1}^{N} q_t(i)\delta_{\ell_t(i)} \in \Delta_K$.
5. Compute the estimated loss of the experts component-wise $g_t(h_t(i))$
6. Update the probability distribution over the experts component-wise

$$q_{t+1}(i) = \frac{\exp\Big(-\eta \sum_{s=1}^{t} g_s\big(h_s(i)\big)\Big)}{\sum_{j=1}^{N} \exp\Big(\eta \sum_{s=1}^{t} g_s\big(h_s(j)\big)\Big)}\,, \qquad \forall 1 \le i \le N\,.$$

Theorem 3. *EXP4 with $\eta = \sqrt{\log N/(KT)}$ satisfies $R_T^{exp} \le 2\sqrt{TK \log N}$.*

Similarly to the variant EXP3.P, we can define a variant EXP4.P to bound the regret with high probability (and thus the expected regret). Furthermore, the above algorithm (and theorem) can be extended to the case where expert advice are distributions $h_t(i) \in \Delta_K$. The algorithm is the same by sampling k_t according to $h_t(i_t)$ and assigning to expert i the loss $\sum_{k=1}^{K} h_t(i)(k) g_t(k)$.

Proof. We can apply the analysis of EXP to a learner using distribution q_t over N actions (here experts) with (full-information) losses $g_t\big(h_t(i)\big)$ for $i \in \{1, \dots, N\}$. We get from Inequality $(*)$

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} q_t(i) \cdot g_t\big(h_t(i)\big) - \min_{1\le j\le N} \sum_{t=1}^{T} g_t\big(h_t(j)\big)\right] \le \eta \sum_{t=1}^{T}\sum_{i=1}^{N} \mathbb{E}\Big[q_t(i) g_t\big(h_t(i)\big)^2\Big] + \frac{\log N}{\eta}\,. \tag{9}$$

Remark that $k_t = h_t(i)$ with probability $q_t(i)$ so that, k_t follows the distribution $p_t \stackrel{\text{def}}{=} \sum_{i=1}^{N} q_t(i)\delta_{h_t(i)}$ knowing the past information $\mathcal{F}_{t-1} \stackrel{\text{def}}{=} \sigma(h_1, i_1, k_1, \dots, i_{t-1}, k_{t-1}, h_t)$. Now, similarly to the proof of EXP3, we compute the expectations. We have for all $k \in [K] \stackrel{\text{def}}{=} \{1, \dots, K\}$

$$\mathbb{E}\Big[g_t(k)\Big|\mathcal{F}_{t-1}\Big] = \mathbb{E}\Big[\frac{\ell_t(k)}{p_t(k)}\mathbb{1}_{\{k=k_t\}}\Big|\mathcal{F}_{t-1}\Big] = \sum_{j=1}^{K} p_t(j)\frac{\ell_t(k)}{p_t(k)}\mathbb{1}_{\{k=j\}} = \ell_t(k)\,,$$

and thus for all $i \in \{1, \dots, N\}$

$$\mathbb{E}\Big[g_t\big(h_t(i)\big)\Big|\mathcal{F}_{t-1}\Big] = \ell_t\big(h_t(i)\big)\,, \tag{10}$$

and

$$\mathbb{E}\left[\sum_{i=1}^{N} q_t(i)\cdot g_t\big(h_t(i)\big)\Big|\mathcal{F}_{t-1}\right] = \sum_{i=1}^{N} q_t(i)\mathbb{E}\Big[g_t\big(h_t(i)\big)\Big|\mathcal{F}_{t-1}\Big] = \sum_{i=1}^{N} q_t(i)\ell_t\big(h_t(i)\big)$$
$$= \mathbb{E}\big[\ell_t\big(h_t(i_t)\big)\big|\mathcal{F}_{t-1}\big] = \mathbb{E}\big[\ell_t(k_t)\big|\mathcal{F}_{t-1}\big]\,. \quad (11)$$

Furthermore,

$$\mathbb{E}\Big[g_t\big(h_t(i)\big)^2\Big|\mathcal{F}_{t-1}\Big] = \sum_{k=1}^{K} p_t(k)\left(\frac{\ell_t\big(h_t(i)\big)}{p_t\big(h_t(i)\big)}\right)^2 \mathbb{1}_{\{k=h_t(i)\}} = \frac{\ell_t\big(h_t(i)\big)^2}{p_t\big(h_t(i)\big)} \le \frac{1}{p_t\big(h_t(i)\big)}\,,$$

and

$$\sum_{i=1}^{N} q_t(i)\mathbb{E}\Big[g_t\big(h_t(i)\big)^2\Big|\mathcal{F}_{t-1}\Big] \le \sum_{i=1}^{N}\frac{q_t(i)}{p_t(h_t(i))} = \mathbb{E}\left[\frac{1}{p_t(h_t(i_t))}\Big|\mathcal{F}_{t-1}\right] = \mathbb{E}\left[\frac{1}{p_t(k_t)}\Big|\mathcal{F}_{t-1}\right] = \sum_{k=1}^{K}\frac{p_t(k)}{p_t(k)} = K\,. \quad (12)$$

Substituting (10), (11), and (12) into Inequality (9) and lower-bounding the expected regret with the pseudo-regret, we get

$$\begin{aligned}
\bar{R}_T^{\exp} &\overset{\text{def}}{=} \max_{1\le i\le N}\mathbb{E}\left[\sum_{t=1}^{T}\ell_t(k_t) - \ell_t\big(h_t(i)\big)\right]\\
&\overset{(10),(11)}{=} \max_{1\le i\le N}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} q_t(i)g_t\big(h_t(i)\big) - g_t\big(h_t(i)\big)\right]\\
&\overset{\text{Jensen}}{\le} \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} q_t(i)g_t\big(h_t(i)\big) - \min_{1\le i\le N} g_t\big(h_t(i)\big)\right]\\
&\overset{(9)}{\le} \eta\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\Big[q_t(i)g_t\big(h_t(i)\big)^2\Big] + \frac{\log N}{\eta}\\
&\overset{(12)}{\le} \eta KT + \frac{\log N}{\eta}\,.
\end{aligned}$$

Optimizing η concludes the proof. □

3 Adversarial Bandits with side information

A natural extension of the previous setting is by adding side (or contextual) information: this is called contextual bandits. It arises in most applications such as recommendation systems or online advertisement. The side information can then be the cookies of a new user to which we need to recommend a product.

Assume that for each time step $t \ge 1$, before doing its prediction k_t the learner observes a context x_t in a finite set $\mathcal{X}$ of contexts. The learner must than learn the best mapping $g : \mathcal{X} \to [K]$ and is evaluated with the contextual pseudo-regret:

$$R_T^{\mathcal{X}} \overset{\text{def}}{=} \max_{g:\mathcal{X}\to[K]}\mathbb{E}\Big[\sum_{t=1}^{T}\ell_t(k_t) - \sum_{t=1}^{T}\ell_t\big(g(x_t)\big)\Big]\,.$$

Similarly, to the stochastic setting, if $\mathcal{X}$ is finite, a simple algorithm consists in running a different copy $EXP3(c)$ of EXP3 for each context $c \in \mathcal{X}$. We denote by $\mathcal{X}$-EXP3 this algorithm. At each time step $t \ge 1$, the learner uses $EXP(x_t)$ to make the prediction. The following theorem follows from Theorem 1.

Theorem 4. *The contextual pseudo-regret of $\mathcal{X}$-EXP3 is upper-bounded as:*

$$R_T^{\mathcal{X}} \leq 2\sqrt{T|\mathcal{X}|K\log K}\,.$$

Proof. Applying the proof of the pseudo-regret bound of EXP3 for each instance $x \in \mathcal{X}$:

$$\max_{j\in[K]} \mathbb{E}\left[\sum_{t=1}^{T} \big(\ell_t(k_t) - \ell_t(j)\big)\mathbb{1}_{\{x_t=x\}}\right] \leq 2\sqrt{K(\log K)T_x}\,,$$

where $T_x = \sum_{t=1}^{T} \mathbb{1}_{\{x_t=x\}}$. Summing over $x \in \mathcal{X}$,

$$\sum_{x\in\mathcal{X}} \max_{j\in[K]} \mathbb{E}\left[\sum_{t=1}^{T} \big(\ell_t(k_t) - \ell_t(j)\big)\mathbb{1}_{\{x_t=x\}}\right] \leq 2\sum_{x\in\mathcal{X}} \sqrt{K(\log K)T_x} \overset{\text{Jensen}}{\leq} 2\sqrt{|\mathcal{X}|K(\log K)T}$$

where the last inequality is by using the concavity of the square root together with $\sum_{s\in\mathcal{X}} T_s = T$. The proof is concluded by remarking that the left-hand side is the contextual pseudo-regret. □

Similarly to the classical lower-bound $O(\sqrt{TK})$, a lower-bound of order $\sqrt{|\mathcal{X}|KT}$ holds under the assumption that a significant proportion of the contexts are used at least a constant fraction of the T rounds. The above bound is nice but the dependency $|\mathcal{X}|$ might be annoying if $\mathcal{X}$ is large.

Exercise 3.1. Generalize the above algorithm and upper-bound when the context-space is continuous and the loss functions are β-Hölder in the contexts.

Competing against the best context set

In some cases, one may want to combine bandit algorithms. For example, we could have in hand different context set $\mathcal{X}$. For each of these sets $\mathcal{X}$, we can bound the pseudo-regret $R_T^{\mathcal{X}}$ using Theorem 4 with $\mathcal{X}$-EXP3 of Section 3, but we would like to find the best set $\mathcal{X}$. To do so, we may want to combine with EXP4 different instances of $\mathcal{X}$-EXP3, each using its own context set $\mathcal{X}$. We can then combine the bounds of Theorem 3 and Theorem 4 to ensure we are competing with the best possible context set $\mathcal{X}$. In this case, each instance of $\mathcal{X}$-EXP3 does not observe their own choice of action but the action chosen by EXP4 which follows a different distribution. The bound of Theorem 3 is valid but the regrets of the experts cannot be bounded using Theorem 4. It is however possible to use a variant of EXP4 to combine bandit algorithms by adding an exploration parameter. We then lose however in the rate of the regret bound which is then of order

$$\max_{\mathcal{X}} R_T^{\mathcal{X}} \leq \mathcal{O}\Big(T^{2/3}\big(\max|\mathcal{X}|K\log K\big)^{1/3}\sqrt{\log M}\Big)$$

where M is the number of context sets $\mathcal{X}$. We refer to Section 4.2.1 of Bubeck et al. [2012] for more details on this application.

References

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.