Sequential learning – Adversarial Bandits

Pierre Gaillard

February 18, 2022

INRIA

Reminder from last weeks

The exponentially weighted average algorithm for bandits

High probability bound on the regret

Adversarial bandits with experts

OGD without Gradients

Setting of an online learning problem/online convex optimization

At each time step $t = 1, \ldots, T$

- the player observes a context $x_t \in \mathcal{X}$ (optional step)
- the player chooses an action $\theta_t \in \Theta$ (compact decision/parameter set);
- the environment chooses a loss function $\ell_t : \Theta \to [0, 1]$;
- 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.

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

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

Previous results

The Exponentially Weighted Average (EWA) forecaster

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

achieves a cumulative regret $R_T \lesssim \sqrt{T \log K}$ when the set of actions is the K-dimensional simplex and for linear losses $\ell_t(p) = p^\top g_t$ with $g_t \in [-1, 1]^K$.

In particular, we saw the intermediate regret-bound if $-\eta g_t(k) \leqslant 1$

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

Note that the loss vectors g_t may depend on past information $p_1, g_1, \ldots, g_{t-1}, p_t$.

This lesson

We will see what we can do with bandit feedback.

At each time step $t = 1, \dots, T$

- the player observes a context $x_t \in \mathcal{X}$ (optional step)
- the player chooses an action $\theta_t \in \Theta$ (compact decision/parameter set);
- the environment chooses a loss function $\ell_t : \Theta \to [0, 1]$;
- 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.

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

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

Reminder from last weeks

The exponentially weighted average algorithm for bandits

High probability bound on the regret

Adversarial bandits with experts

OGD without Gradients

Adversarial multi-armed bandit and pseudo-regret

Setting: $\Theta = \{1, \ldots, K\}$. At round t, the player chooses an action $k_t \in \{1, \ldots, K\}$ and suffers and observes the loss $\ell_t(k_t) \in [0, 1]$ only.

Regret with respect to action $k \in [K]$ by

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

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.

Definition (Pseudo-regret)

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

Oblivious vs adaptive adversary

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

The expectation is 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 $\ell_1, \ldots, \ell_t$ are chosen in advance before the game starts and do not depend on the past player decisions $k_1, \ldots, k_T$. 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 $t \geqslant 1$ may depend on past information $\sigma(k_1, \ldots, k_{t-1})$. It is thus random. By Jensen's inequality $\max_{k \in [K]} \mathbb{E}\big[R_T(k)\big] \leqslant \mathbb{E}\big[\max_{k \in [K]} R_T(k)\big]$ and thus $\bar{R}_T \leqslant \mathbb{E}[R_T]$.

How to use EWA for bandits?

The Exponentially Weighted Average (EWA) forecaster

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

Question: Can we use directly $p_t(k)$ as defined by EWA with $g_t = (\ell_t(1), \ldots, \ell_t(K))$ and sample $k_t \sim p_t$ as we did for random EWA?

☐ Yes ☐ No

How to use EWA for bandits?

The Exponentially Weighted Average (EWA) forecaster

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

Question: Can we use directly $p_t(k)$ as defined by EWA with $g_t = (\ell_t(1), \dots, \ell_t(K))$ and sample $k_t \sim p_t$ as we did for random EWA?

Answer: No, since the player does not observe $\ell_t(k)$ for $k \neq k_t$ and cannot compute p_t.

How to use EWA for bandits?

The Exponentially Weighted Average (EWA) forecaster

$$\boxed{p_t(k) = \frac{e^{-\eta \sum_{s=1}^{t-1} g_s(k)}}{\sum_{j=1}^{K} e^{-\eta \sum_{s=1}^{t-1} g_s(j)}}} \qquad \text{(EWA)}$$

Question: What about setting using $\ell_t(k)$ if we observe it and 0 otherwise:

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

☐ Yes ☐ No

How to use EWA for bandits?

The Exponentially Weighted Average (EWA) forecaster

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

Question: What about setting using $\ell_t(k)$ if we observe it and 0 otherwise:

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

Answer: No, because this estimate would be 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.

How to use EWA for bandits?

The Exponentially Weighted Average (EWA) forecaster

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

Therefore, we choose

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

which leads to the algorithm EXP3 detailed below.

Exponential Weights for bandits

EXP3

Parameter: $\eta > 0$

Initialize: $p_1 = \left(\frac{1}{K}, \ldots, \frac{1}{K}\right)$

For $t = 1, \ldots, T$

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

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

Pseudo-Regret bound for EXP3

$$p_{t+1}(k) = \frac{e^{-\eta \sum_{s=1}^{t} g_s(k)}}{\sum_{j=1}^{K} e^{-\eta \sum_{s=1}^{t} g_s(j)}}, \quad \text{where } g_s(k) = \frac{\ell_s(k)}{p_s(k)} \mathbb{1}_{\{k=k_s\}} \tag{EXP3}$$

Theorem 1

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

$$\bar{R}_T \stackrel{\text{def}}{=} \max_{k \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(k_t) - \sum_{t=1}^{T} \ell_t(k)\right] \leqslant 2\sqrt{KT \log K}\,.$$

Proof

Applying EWA to the estimated losses $g_t(j)$ that are completely observed and taking the expectation:

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

The rest of the proof consists in computing the expectations:

$$\mathbb{E}\left[p_t \cdot g_t\right] = \mathbb{E}\left[\ell_t(k_t)\right], \qquad \mathbb{E}\left[g_t(j)\right] = \mathbb{E}\left[\ell_t(j)\right] \qquad \text{and} \qquad \mathbb{E}\left[p_t \cdot g_t^2\right] \leqslant K \qquad (1)$$

You have 15 min to try to prove some of these equations.

Proof

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.

1) Proof that $\mathbb{E}\big[p_t \cdot g_t\big] = \mathbb{E}\big[\ell_t(k_t)\big]$

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

2) Proof that $\mathbb{E}\big[g_t(j)\big] = \mathbb{E}\big[\ell_t(j)\big]$

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

Proof

Therefore, using

$$\mathbb{E}\left[p_t \cdot g_t\right] = \mathbb{E}\left[\ell_t(k_t)\right] \qquad \text{and} \qquad \mathbb{E}\left[g_t(j)\right] = \mathbb{E}\left[\ell_t(j)\right] \tag{2}$$

we have

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

Proof

3) Proof that $\mathbb{E}\left[p_t \cdot g_t^2\right] \leqslant K$

$$\mathbb{E}\left[p_t \cdot g_t^2\right] = \mathbb{E}\left[\sum_{j=1}^{K} p_t(j) g_t(j)^2\right] = \mathbb{E}\left[\sum_{j=1}^{K} p_t(j)\, \mathbb{E}\left[g_t(j)^2 \,\middle|\, \mathcal{F}_{t-1}\right]\right]$$

$$= \mathbb{E}\left[\sum_{j=1}^{K} \sum_{k=1}^{K} p_t(j) p_t(k) \left(\frac{\ell_t(j)}{p_t(j)} \mathbb{1}_{\{j=k\}}\right)^2\right]$$

$$= \mathbb{E}\left[\sum_{j=1}^{K} \sum_{k=1}^{K} p_t(k) \frac{\ell_t(j)^2}{p_t(j)} \mathbb{1}_{\{j=k\}}\right]$$

$$= \mathbb{E}\left[\sum_{j=1}^{K} \ell_t(j)^2\right] \leqslant K\,.$$

4) Conclusion. Substituting into Inequality $(*)$ yields

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

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

Limit of the result

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

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

but not

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

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

Reminder from last weeks

The exponentially weighted average algorithm for bandits

High probability bound on the regret

Adversarial bandits with experts

OGD without Gradients

Gains versus losses

We switch the analysis from losses $\ell_t(k)$ to rewards $r_t(k) = 1 - \ell_t(k) \in [0,1]$.

Remark that the loss and gain versions are symmetric via the transformation $r_t(k) = 1 - \ell_t(k)$. The regret in terms of gains is defined as

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

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

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

High-level idea of EXP3.P

The high-level idea of the next algorithm is to ensure that the estimators $g_t(k)$ of the gains $r_t(k)$ satisfy

$$\mathbb{E}\left[\max_j \sum_{t=1}^{T} g_t(j)\right] \geqslant \mathbb{E}\left[\max_j \sum_{t=1}^{T} r_t(j)\right] \tag{5}$$

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).

To ensure (5), we add a bias term β to the estimators $g_t(k)$ as follows:

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

Estimator of EXP3.P

$$g_t(k) \stackrel{\text{def}}{=} \frac{r_t(k)\mathbb{1}_{\{k=k_t\}} + \beta}{p_t(k)}$$

The estimator is indeed biased

$$\mathbb{E}\left[g_t(k) \middle| \mathcal{F}_{t-1}\right] = r_t(k) + \frac{\beta}{p_t(k)},$$

Lemma 1

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

$$\sum_{t=1}^{T} g_t(j) \geqslant \sum_{t=1}^{T} r_t(j) - \frac{\log(1/\delta)}{\beta}.$$

Proof of the lemma

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

$$\begin{aligned}\mathbb{P}\left(\sum_{t=1}^{T} g_t(j) \geqslant \sum_{t=1}^{T} r_t(j) - \frac{\log(1/\delta)}{\beta}\right) &= \mathbb{P}\left(\exp\left(\beta \sum_{t=1}^{T}\left(r_t(j) - g_t(j)\right)\right) \geqslant \delta^{-1}\right) \\ &\leqslant \delta \mathbb{E}\left[\exp\left(\beta \sum_{t=1}^{T}\left(r_t(j) - g_t(j)\right)\right)\right].\end{aligned}$$

It only remains to upper-bound the expectation in the right-hand side by 1, which we do now.

Proof

Since $\beta \in (0,1)$ and $g_t(j) \geqslant \beta/p_t(j)$, we have $\beta(r_t(j) - g_t(j) + \beta/p_t(j)) \leqslant 1$. Therefore, we can use the inequality $e^x \leqslant 1 + x + x^2$ for $x \leqslant 1$, which entails

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\beta\left(r_t(j) - g_t(j)\right)\right)\middle|\mathcal{F}_{t-1}\right] \\
&\quad = \mathbb{E}\left[\exp\left(\beta\left(r_t(j) - g_t(j) + \frac{\beta}{p_t(j)}\right)\right)\middle|\mathcal{F}_{t-1}\right]\exp\left(-\frac{\beta^2}{p_t(j)}\right) \\
&\quad \leqslant \mathbb{E}\left[\left(1 + \beta\cancel{\left(r_t(j) - g_t(j) + \frac{\beta}{p_t(j)}\right)} + \beta^2\left(r_t(j) - g_t(j) + \frac{\beta}{p_t(j)}\right)^2\right)\middle|\mathcal{F}_{t-1}\right]e^{-\frac{\beta^2}{p_t(j)}} \\
&\quad = \left(1 + \beta^2\mathbb{E}\left[\left(r_t(j) - g_t(j) + \frac{\beta}{p_t(j)}\right)^2\middle|\mathcal{F}_{t-1}\right]\right)e^{-\frac{\beta^2}{p_t(j)}}
\end{aligned}$$

where the last equality is because $\mathbb{E}\left[g_t(k)\middle|\mathcal{F}_{t-1}\right] = r_t(k) + \frac{\beta}{p_t(k)}$ and because $p_t(j)$ is $\mathcal{F}_{t-1}$-measurable.

Proof

Now,

$$\mathbb{E}\left[\left(r_t(j) - g_t(j) + \frac{\beta}{p_t(j)}\right)^2 \middle| \mathcal{F}_{t-1}\right] = \mathrm{Var}\left(g_t(j) \middle| \mathcal{F}_{t-1}\right) = \mathrm{Var}\left(\frac{r_t(j)\mathbb{1}_{\{j=k_t\}}}{p_t(j)} \middle| \mathcal{F}_{t-1}\right)$$

$$\leqslant \mathbb{E}\left[\left(\frac{r_t(j)\mathbb{1}_{\{j=k_t\}}}{p_t(j)}\right)^2 \middle| \mathcal{F}_{t-1}\right] \leqslant \mathbb{E}\left[\frac{\mathbb{1}_{\{j=k_t\}}}{p_t(j)^2} \middle| \mathcal{F}_{t-1}\right] = \sum_{k=1}^{K} \frac{p_t(k)\mathbb{1}_{\{j=k\}}}{p_t(j)^2} = \frac{1}{p_t(j)} .$$

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

$$\mathbb{E}\left[\exp\left(\beta\left(r_t(j) - g_t(j)\right)\right) \middle| \mathcal{F}_{t-1}\right] \leqslant \left(1 + \frac{\beta^2}{p_t(j)}\right) e^{-\beta^2/p_t(j)} \leqslant 1 .$$

The proof is concluded by induction

$$\mathbb{E}\left[\exp\left(\beta \sum_{t=1}^{T} \left(r_t(j) - g_t(j)\right)\right)\right] = \mathbb{E}\left[\underbrace{\mathbb{E}\left[\exp\left(\beta\left(r_T(j) - g_T(j)\right)\right) \middle| \mathcal{F}_{T-1}\right]}_{\leqslant 1} \exp\left(\beta \sum_{t=1}^{T-1} \left(r_t(j) - g_t(j)\right)\right)\right]$$

$$\leqslant \mathbb{E}\left[\exp\left(\beta \sum_{t=1}^{T-1} \left(r_t(j) - g_t(j)\right)\right)\right] \leqslant \ldots \leqslant 1 .$$

EXP3.P

EXP3.P

Parameters: $\eta > 0, \beta \in (0,1), \gamma \in (0,1)$
Initialize: $p_1 = \left(\frac{1}{K}, \dots, \frac{1}{K}\right)$

For $t = 1, \dots, T$

- draw $k_t \sim p_t$; receive reward $r_t(k_t) = 1 - \ell_t(k_t)$ and observe $r_t(k_t) \in [0,1]$;
- update for all $k \in \{1, \dots, K\}$

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

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

The weights $p_t(k)$ of EXP3.P are necessary larger than γ/K and thus $|\eta g_t(j)| \leqslant 1$ as soon as $\eta(1+\beta)K/\gamma \leqslant 1$.

Regret bound for Exp3.P

Theorem 2

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

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

with probability at least $1-\delta$.

With the choice $\delta = 1/T$ it yields

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

Proof of Theorem 2

Step 1. Apply the classical bound of EWA Defining the weights that would assign EXP3,

$$q_t(j) \stackrel{\text{def}}{=} \frac{e^{\eta \sum_{s=1}^{t-1} g_s(j)}}{\sum_{k=1}^{K} e^{\eta \sum_{s=1}^{t-1} g_s(k)}} ,$$

we have $p_t \stackrel{\text{def}}{=} (1-\gamma) q_t + \gamma / K$.

we get from Inequality (4) applied with $g_t(j)$,

$$\max_{j \in [K]} \sum_{t=1}^{T} g_t(j) \leqslant \sum_{t=1}^{T} q_t \cdot g_t + \eta \sum_{t=1}^{T} q_t \cdot g_t^2 + \frac{\log K}{\eta} .$$

where we used $\eta g_t(j) \leqslant 1$ because $\eta(1+\beta)K/\gamma \leqslant 1$.

Proof

Step 2. Rewrite the bound with p_t instead of q_t Now, we use that $p_t \stackrel{\text{def}}{=} (1-\gamma)q_t + \gamma/K$, which entails $q_t = (p_t - \gamma/K)/(1-\gamma) \leqslant p_t/(1-\gamma)$. Substituting into the above inequality

$$(1-\gamma)\max_{j\in[K]}\sum_{t=1}^{T} g_t(j) \leqslant \sum_{t=1}^{T} p_t \cdot g_t + \eta \sum_{t=1}^{T} p_t \cdot g_t^2 + \frac{\log K}{\eta}\,. \tag{7}$$

Step 3. Replace g_t with r_t. By definition of g_t,

$$p_t \cdot g_t = \sum_{j=1}^{K} p_t(j) g_t(j) = \sum_{j=1}^{K} \left(r_t(j)\mathbb{1}_{\{j=k_t\}} + \beta\right) = r_t(k_t) + K\beta\,.$$

and since $p_t(j)g_t(j) \leqslant (1+\beta)$,

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

Therefore, substituting into Inequality (7) gives

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

Proof

We had

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

Reorganizing, we get

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

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

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

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

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

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

Reminder from last weeks

The exponentially weighted average algorithm for bandits

High probability bound on the regret

Adversarial bandits with experts

OGD without Gradients

Setting of adversarial bandits with experts

Setting

At each time step $t = 1, \ldots, T$

- N experts propose recommendations $h_t(i) \in [K]$ for $i \in [N]$
- the environment chooses a loss function $\ell_t : \Theta \to [0, 1]$;
- the player chooses an action $k_t \in [K]$
- the player suffers loss $\ell_t(k_t)$
- the player observes the loss of the chosen action only: $\ell_t(k_t)$

Goal: compete with the best expert, i.e., minimize

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

with respect to the experts.

EXP3 solution

By using EXP3 on the set of experts instead of the set of actions, we would get

$$\bar{R}_T \leqslant \sqrt{TN \log N}\,.$$

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

EXP4

Parameter: $\eta > 0$ Initialize: $q_1 = \left(\frac{1}{N}, \ldots, \frac{1}{N}\right)$.

For each round $t = 1, \ldots, n$

1. Get expert advice $h_t(1), \ldots, h_t(N) \in [K]$
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

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

Regret of EXP4

Theorem 3

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

Proof left as exercise.

Reminder from last weeks

The exponentially weighted average algorithm for bandits

High probability bound on the regret

Adversarial bandits with experts

OGD without Gradients

Beyond finite set of actions?

At each time step $t = 1, \ldots, T$

- the player observes a context $x_t \in \mathcal{X}$ (optional step)
- the player chooses an action $\theta_t \in \Theta$ (compact decision/parameter set);
- the environment chooses a loss function $\ell_t : \Theta \to [0, 1]$;
- 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.

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

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

This lecture: we saw variants of EXP3 when Θ is finite.
What if the losses ℓ_t are convex but Θ is any bounded convex set in $\mathbb{R}^d$?

Online Gradient Descent

In the full information setting (when gradient can be observed), we saw OGD algorithm:

$$\boxed{\theta_{t+1} \leftarrow \mathrm{Proj}_\Theta\left(\theta_t - \eta \nabla \ell_t(\theta_t)\right)}$$

Theorem 4 (Regret of OGD)

Let $D, G, \eta > 0$. Assume that Θ has diameter bounded by D and the convex losses have sub-Gradients bounded by G in ℓ_2-norm, the regret of OGD satisfies

$$\sum_{t=1}^{T} \ell_t(\theta_t) - \min_{\theta \in \Theta} \sum_{t=1}^{T} \ell_t(\theta) \leqslant DG\sqrt{T}.$$

How to adapt this algorithm to the bandit setting? That is, when only $\ell_t(\theta_t)$ are observed and not $\nabla \ell_t(\theta_t)$?

Point-wise gradient estimators

$$\boxed{\theta_{t+1} \leftarrow \mathrm{Proj}_\Theta\left(\theta_t - \eta \nabla \ell_t(\theta_t)\right)}$$

Similarly to EXP3, the idea is to replace the gradient in OGD with unbiased estimators. That is try to find an observable random variable $\widehat{g}_t$ that satisfies

$$\mathbb{E}[\widehat{g}_t] \approx \nabla \ell_t(\theta_t)$$

Example: one-dimensional gradient estimate

$$\ell'(x) = \lim_{\delta \to 0} \frac{\ell(x+\delta) - \ell(x-\delta)}{2\delta} .$$

Thus we can define

$$\widehat{g}(x) = \begin{cases} \frac{\ell(x+\delta)}{\delta} & \text{with proba } \frac{1}{2} \\ -\frac{\ell(x-\delta)}{\delta} & \text{with proba } \frac{1}{2} \end{cases} \qquad \text{which yields} \quad \mathbb{E}[\widehat{g}(x)] = \frac{\ell(x+\delta) - \ell(x-\delta)}{2\delta} .$$

Thus in expectation, for small δ, $\widehat{g}(x)$ approximates $\ell'(x)$.

Point-wise gradient estimators: multi-dimensional case

We show here how the one-dimensional pointwise gradient estimator can be extended to the multi-dimensional case.

We define $\widehat{\ell}_t$ to be a smoothed version of the loss:

$$\widehat{\ell}_t(\theta) = \mathbb{E}_v \left[\ell_t(\theta + \delta v)\right]$$

where $v \sim \mathit{Unif}(\mathbb{B})$. If δ is small, $\widehat{\ell}_t$ is a good approximation of ℓ_t.

Lemma 2

Let $\widehat{\ell}_t(\theta) = \mathbb{E}\left[\ell_t(\theta + \delta v)\right]$ where $v \sim \mathit{Unif}(\mathbb{B})$ be a smoothed version of the loss, then

$$\mathbb{E}_u\left[\frac{d}{\delta}\ell_t(\theta_t + \delta u)u\right] = \nabla\widehat{\ell}_t(\theta)\,.$$

Proof.

Left as exercise. See Lem. 6.7, Hazan et al., "Introduction to online convex optimization", 2016. □

OGD without Gradients

Similarly to EXP3, the idea is to replace the gradient in OGD with unbiased estimators.

OGD without gradients

For $t = 1, \dots, T$

- Draw $u_t \in \mathbb{S}$ uniformly at random in the unit sphere
- Set $\widehat{\theta}_t = \theta_t + \delta u_t$ a random perturbation of the current point θ_t
- Play $\widehat{\theta}_t$
- Estimate the gradient in θ_t with

$$\widehat{g}_t = \frac{d}{\delta} \ell_t(\widehat{\theta}_t) u_t$$

- Update

$$\theta_{t+1} \leftarrow \mathrm{Proj}_{\Theta_\delta}\left(\theta_t - \eta \widehat{g}_t\right)$$

where $\Theta_\delta = \left\{\theta \in \Theta \quad \text{s.t} \quad \theta + \delta u \in \Theta \quad \forall u \in \mathbb{S}\right\}$

Regret of OGD without gradients

OGD without gradients:

$$\theta_{t+1} \leftarrow \mathrm{Proj}_{\Theta_\delta}(\theta_t - \eta \widehat{g}_t) \qquad \text{where} \quad \widehat{g}_t = \frac{d}{\eta} \ell_t(\widehat{\theta}_t) u_t \text{ and } \widehat{\theta}_t = \theta_t + \delta u_t$$

Theorem 5

If the losses are in $[-1, 1]$ *and* G*-Lipschitz, OGD without gradients with parameters* $\delta = \min\{D, (1/2)\sqrt{Dd/G}\,T^{-1/4}\}$ *and* $\eta = D\delta/(dT^{1/2})$ *satisfies the expected regret bound*

$$\sum_{t=1}^{T} \mathbb{E}[\ell_t(\widehat{\theta}_t)] - \min_{\theta \in \Theta} \sum_{t=1}^{T} \ell_t(\theta) \leqslant 2d\sqrt{T} + 2\sqrt{GDd}\,T^{3/4}.$$

Proof (Step 1)

Denote

$$\theta^* \in \arg\min_{\theta \in \Theta} \sum_{t=1}^{T} \ell_t(\theta) \qquad \text{and} \qquad \theta^*_\delta = \mathsf{Proj}_{\Theta_\delta}(\theta^*)\,.$$

Then,

$$\left\|\theta^* - \theta^*_\delta\right\| \leqslant \delta$$

Thus, if the losses are G-Lipschitz

$$\begin{aligned} R_T := \sum_{t=1}^{T} \mathbb{E}\big[\ell_t(\widehat{\theta}_t)\big] - \sum_{t=1}^{T} \ell_t(\theta^*) &\leqslant \sum_{t=1}^{T} \mathbb{E}\big[\ell_t(\widehat{\theta}_t)\big] - \sum_{t=1}^{T} \ell_t(\theta^*_\delta) + \delta TG \\ &\leqslant \sum_{t=1}^{T} \mathbb{E}\big[\ell_t(\theta_t)\big] - \sum_{t=1}^{T} \ell_t(\theta^*_\delta) + 2\delta TG \\ &\leqslant \sum_{t=1}^{T} \mathbb{E}\big[\widehat{\ell}_t(\theta_t)\big] - \sum_{t=1}^{T} \widehat{\ell}_t(\theta^*_\delta) + 4\delta TG \qquad (*) \end{aligned}$$

where $\widehat{\ell}_t(\theta) = \mathbb{E}_v\left[\ell_t(\theta + \delta v)\right]$ with $v \sim \mathit{Unif}(\mathbb{B})$ are the smoothed versions of the losses.

Proof (Step 2)

Now, recall that the algorithm runs OGD with $\widehat{g}_t$ in place of the gradients:

$$\theta_{t+1} \leftarrow \mathrm{Proj}_{\Theta_\delta}\left(\theta_t - \eta \widehat{g}_t\right)$$

Defining the pseudo-loss $h_t(\theta) = \widehat{\ell}_t(\theta) + (\widehat{g}_t - \nabla \widehat{\ell}_t(\theta_t))^\top \theta$, we can see that

$$\nabla h_t(\theta_t) = \nabla \widehat{\ell}_t(\theta_t) + \widehat{g}_t - \nabla \widehat{\ell}_t(\theta_t) = \widehat{g}_t .$$

Therefore, the algorithm actually runs OGD on the losses h_t and thus satisfies the OGD regret bound (see Lecture 2)

$$\sum_{t=1}^{T} h_t(\theta_t) - \sum_{t=1}^{T} h_t(\theta_\delta^*) \leqslant \frac{D^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|g_t\|^2 .$$

Furthermore, by construction of the gradient estimator, we have $\mathbb{E}_{u_t}\left[\widehat{g}_t\right] = \nabla \widehat{\ell}_t(\theta_t)$, which yields

$$\mathbb{E}_{u_t}\left[h_t(\theta_t)\right] = \widehat{\ell}_t(\theta_t) \quad \text{and} \quad \mathbb{E}_{u_t}\left[h_t(\theta_\delta^*)\right] = \widehat{\ell}_t(\theta_\delta^*)$$

Thus taking the expectation in the previous regret bound entails

$$\sum_{t=1}^{T} \mathbb{E}[\widehat{\ell}_t(\theta_t)] - \sum_{t=1}^{T} \widehat{\ell}_t(\theta_\delta^*) = \mathbb{E}\left[\sum_{t=1}^{T} h_t(\theta_t) - \sum_{t=1}^{T} h_t(\theta_\delta^*)\right] \leqslant \frac{D^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \mathbb{E}\left[\|g_t\|^2\right] \qquad (**)$$

Proof (Step 3)

Combining the two bounds (*) and (**) that we have proved, we get

$$R_T \leqslant \frac{D^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \mathbb{E}\left[\|g_t\|^2\right] + 4\delta TG$$

Then, since $|\ell_t(\theta)| \leqslant 1$ for all $\theta \in \Theta$,

$$\|g_t\|^2 = \left(\frac{d}{\delta}\ell_t(\widehat{\theta}_t)\right)^2 \leqslant \frac{d^2}{\delta^2}$$

This finally yields the regret

$$R_T \leqslant \frac{D^2}{2\eta} + \frac{\eta d^2 T}{2\delta^2} + 4\delta TG \leqslant 2d\sqrt{T} + 2\sqrt{GDd}\,T^{3/4}$$

for the choices of δ and η.

More on convex bandits

Convex bandits is still an active research area with many open problems.

The above regret bound of order $O(T^{3/4})$ is suboptimal.

More complicated methods can achieve $O(\sqrt{T})$ regret but with sub-optimal dependence on d and worst computational complexities.

More information can be found in Hazan et al., "Introduction to online convex optimization", 2016.

References

Thank you!

- Cesa-Bianchi, Nicolo and Gábor Lugosi. Prediction, learning, and games. Cambridge university press, 2006.
- Hazan, Elad et al. "Introduction to online convex optimization". In: Foundations and Trends® in Optimization 2.3-4 (2016), pp. 157–325.
- Lattimore, Tor and Csaba Szepesvári. "Bandit algorithms". In: preprint (2019).
- Shalev-Shwartz, Shai et al. "Online learning and online convex optimization". In: Foundations and Trends® in Machine Learning 4.2 (2012), pp. 107–194.