CS639: Algorithmic Game Theory & Learning

Chapter 4: Introduction to Online Learning

Kirthevasan Kandasamy

UW-Madison

Outline

1. Learning from Expert Advice
2. The Hedge algorithm
3. Adversarial bandits and the EXP3 algorithm.

Slides are intended as teaching aids only and do not include all material discussed in class. Students are strongly encouraged to attend lectures and take their own notes.

Ch 4.1: Learning from Expert Advice

Example: Route picking (Example 18.3.1 in KP). On each day, you will choose from one of K routes to campus. Your goal is to minimize your travel time, which depends on the traffic. The traffic on each route can change on each day, and you do not know this ahead of time.

- If you choose route $i \in [K]$ on day t, you incur latency (loss) $\ell_t(i)$.

- At the end of the day, you observe latencies for *all* routes: $\ell_t = (\ell_t(1), \ldots, \ell_t(K))$.

- On day t, you can use the observed latency on all previous $t - 1$ days to determine which route to take on that day.

Question: How would you choose a route on day t based on past information?

Learning with expert advice, a.k.a the “Experts problem”

Problem set up: There are a set of K actions (experts), denoted $[K]$.
- On round t, a *learner* chooses a probability distribution $p_t = (p_t(1), \ldots, p_t(K)) \in \Delta_K$. An action A_t is sampled from p_t and played.
- An *adversary* (environment) simultaneously (without knowledge of A_t) chooses a loss vector $\ell_t = (\ell_t(1), \ldots, \ell_t(K)) \in [0,1]^K$, where $\ell_t(i)$ is the loss for action i.
- The learner incurs expected loss $\mathbb{E}[\ell_t(A_t)] = \sum_{i=1}^K p_t(i)\ell_t(i) = p_t^\top \ell_t$.
- The learner observes the entire loss vector ℓ_t, i.e the losses for *all* actions.

A learner is characterized by a *policy (algorithm)* π, which maps the history to a distribution over $[K]$, *i.e.*, $p_t = \pi(\{(p_s, A_s, \ell_s)\}_{s=1}^{t-1})$.

Measuring the performance of a policy via regret

The expected cumulative loss of a learner after T rounds is $\sum_{t=1}^{T} p_t^\top \ell_t$.
We want this loss to be small, but we have to measure against a meaningful baseline. If all actions incur large losses, then even the best policy will have large loss. If all actions incur small losses, then even poor policies will appear to perform well.

Regret. The regret of a policy π against a sequence of losses $\ell = (\ell_1, \dots, \ell_T) \in [0,1]^{K \times T}$ is the difference between the cumulative loss incurred by the policy and the cumulative loss of the single best probability distribution in hindsight,

$$R_T(\pi, \ell) = \sum_{t=1}^{T} p_t^\top \ell_t \; - \; \min_{p \in \Delta_K} \sum_{t=1}^{T} p^\top \ell_t. \tag{1}$$

But we can also write, (Why?)

$$R_T(\pi, \ell) = \sum_{t=1}^{T} p_t^\top \ell_t \; - \; \min_{i \in [K]} \sum_{t=1}^{T} \ell_t(i). \tag{2}$$

Measuring the performance of a policy via regret (cont'd)

$$R_T(\pi, \ell) = \sum_{t=1}^{T} p_t^\top \ell_t \; - \; \min_{p \in \Delta_K} \sum_{t=1}^{T} p^\top \ell_t = \sum_{t=1}^{T} p_t^\top \ell_t \; - \; \min_{i \in [K]} \sum_{t=1}^{T} \ell_t(i).$$

We wish to design a policy π to achieve small regret. In particular, we want the regret R_T to be sublinear in T,

$$R_T(\pi, \ell) \in o(T) \qquad \Longleftrightarrow \qquad \lim_{T \to \infty} \frac{1}{T} R_T(\pi, \ell) = 0.$$

We call a policy which satisfies the above condition a *no-regret policy*.

Oblivious vs Adaptive Adversaries *(read at home)*

$$R_T(\pi, \ell) = \sum_{t=1}^{T} p_t^\top \ell_t \; - \; \min_{p \in \Delta_K} \sum_{t=1}^{T} p^\top \ell_t = \sum_{t=1}^{T} p_t^\top \ell_t \; - \; \min_{i \in [K]} \sum_{t=1}^{T} \ell_t(i).$$

For the interested reader only: In online learning, we draw distinctions between two types of adversaries (environments).

1. An **oblivious adversary** chooses the entire loss sequence $\ell = (\ell_1, \ldots, \ell_T)$ ahead of time (*i.e.*, before round 1), possibly with knowledge of the learner's policy.
2. An **adaptive adversary** can choose the losses $\ell = (\ell_1, \ldots, \ell_T)$ on the fly, *i.e.*, can choose ℓ_t based on algorithm's behavior in previous $t-1$ rounds.

The proofs in this chapter apply to both oblivious and adaptive adversaries. However, the above regret is typically meaningful only under an oblivious adversary, since it compares a policy's performance on a *fixed* loss sequence. For adaptive adversaries, one usually studies a stronger notion of regret that models the adversary itself as a policy.

That said, there are a handful cases where this regret remains meaningful under an adaptive adversary. Analyzing no-regret dynamics in games happens to be one of them!

Why study online learning?

- Immediately (in Chapter 5), we will use tools from online learning to:
  - Prove the minimax theorem.
  - Study if *no-regret dynamics* converges to equilibrium behavior in games.
  - Build on this to develop algorithms to approximate NE in zero sum games, and CE/CCE in general sum games.
- In the future (in Chapter 7), we will use online learning to learn unknowns in mechanism design.

Viewing the learner as a player in a game

Recall the congestion game example from Chapter 1:

There are n players and a road network (graph) with edge set E. Player i wishes to drive from point s_i to t_i. Player i's action space $\mathcal{A}_i$ consists of all paths from s_i to t_i. Let $a_i \in \mathcal{A}_i$ be i's action and $a = \{a_i\}_{i=1}^n$ be a given action profile. Let $N_e(a) = |\{i \in [n]; e \in a\}|$ be the number of drivers using edge e. The latency to cross edge e when there are N drivers on the edge is $\ell_e(N)$, where $\ell_e : \mathbb{N} \to \mathbb{R}_+$ is an edge-specific nondecreasing function. Driver i's total latency (cost/loss) under an action profile a is, $c_i(a) \triangleq \sum_{e \in a_i} \ell_e(N_e(a))$.

For a player i, choosing an optimal action requires knowledge of others' strategies, which she may not know.

Motivating questions. Suppose this game is being repeated over multiple days, with each player choosing their action on each day via some policy:

- **Ch4.** Can we design policies so as to guarantee no regret (with loss = latency)?
- **Ch5.** If all players adopt such behavior (called no-regret dynamics), does it converge to an equilibrium?

Ch 4.2: The Hedge algorithm

How do we design a policy for the experts problem?

- On round t, the learner chooses a distribution $p_t \in \Delta_K$.
- The learner observes the entire loss vector ℓ_t, i.e losses for *all* actions.

The most straightforward idea is called "Follow the leader (FTL)", where you pick the action which appears to be the best in the current round.

Follow the leader (FTL).

$$\text{Choose } A_t = \operatorname*{argmin}_{a \in [K]} \sum_{s=1}^{t-1} \ell_s(a),$$

And then set $p_t(A_t) = 1$ and $p_t(j) = 0$ for all $j \neq A_t$.

Failure cases for FTL

Consider $K = 2$ actions, with the following losses:

$$\ell_t = \begin{cases} (0.5, 0) & \text{if } t = 1, \\ (1, 0) & \text{if } t \text{ is odd}, \\ (0, 1) & \text{if } t \text{ is even}. \end{cases}$$

Then, FTL will choose $A_t = 1$ on even t and $A_t = 2$ on odd $t \geq 1$.

Total loss of FTL is at least $T - 1$.
The best loss in hindsight will have loss at most $T/2$.
Therefore, the regret is least $T/2 - 1 \in \Omega(T)$.

Extend this example to show that *any* deterministic policy will fail. Recall that we only need to find one bad loss vector. *(try at home)*

The Hedge algorithm

Main idea: Replace the minimum in FTL with a randomized “soft” minimum.

Algorithm: Hedge (a.k.a exponential weights)

```
▶ Given: time horizon $T$, learning rate $\eta$.          # Will specify $\eta$ later.
▶ Let $L_0 = \mathbf{0}_K$.          # $L_t \in \mathbb{R}_+^K$ will maintain cumulative losses.
▶ for $t = 1, \ldots, T$,
    ▶ Construct $p_t \in \Delta([K])$ as follows,          # $\pi(\cdot|\mathtt{history}) = p_t(\cdot)$.
          $p_t(a) \leftarrow \frac{e^{-\eta L_{t-1}(a)}}{\sum_{j=1}^{K} e^{-\eta L_{t-1}(j)}}$
    ▶ Observe $\ell_t$. Update $L_t(a) \leftarrow L_{t-1}(a) + \ell_t(a)$ for all $a \in [K]$.
```

Intuition. If an action has large losses, it gets discounted more in L_t and has a smaller probability of getting selected in the future.

Regret analysis for the experts problem: preparation

For a given sequence of probability distributions $p = (p_1, \dots, p_T)$, define the pseudo-regret relative to an action $a \in [K]$,

$$\overline{R}_T(p, \ell, a) \triangleq \sum_{t=1}^{T} p_t^\top \ell_t - \sum_{t=1}^{T} \ell_t(a),$$

If we can bound $\overline{R}_T(p, \ell, a)$ for all actions $a \in [K]$, we can bound the regret for Hedge: for any policy π which chooses p_t as a function of the observed losses $\{\ell_s\}_{s=1}^{t-1}$,

$$R_T(\pi, \ell) = \sum_{t=1}^{T} p_t^\top \ell_t \; - \; \min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a) = \overline{R}_T\big(\, p(\pi, \ell),\, \ell,\, a^{\mathrm{OPT}}(\ell)\,\big)$$

where, $a^{\mathrm{OPT}}(\ell) = \mathrm{argmin}_{a \in [K]} \sum_t \ell_t(a)$ is the best action in hindsight for the given loss sequence, and $p(\pi, \ell)$ is the sequence of probability vectors generated when π interacts with ℓ.

Regret analysis of Hedge: main technical lemma

For a given $\ell_t \in \mathbb{R}_+^K$, define ℓ_t^2 so that $\ell_t^2(i) = (\ell_t(i))^2$.

Lemma. Let $p = (p_1, \ldots, p_T)$ be the sequence of probability vectors chosen by Hedge with learning rate $\eta \in [0, 1]$. Then, for any set of loss vectors $\ell = (\ell_1, \ldots, \ell_T)$, where $\ell_t \in \mathbb{R}_+^K$ and any $a \in [K]$, if $p_t^\top \ell_t \leq 1$ for all t, we have

$$\overline{R}_T(p, \ell, a) \leq \frac{\log(K)}{\eta} + \eta \sum_{t=1}^{T} p_t^\top \ell_t^2.$$

We are presenting a more general (than immediately necessary) version of this lemma, since we will build on Hedge when studying adversarial bandits.

Let us first use this lemma to bound the regret of Hedge. We will prove this lemma immediately after.

Regret bound of Hedge (using previous lemma)

Theorem. Suppose $\ell_t \in [0,1]^K$ for all t, and we choose $\eta = \sqrt{\frac{\log(K)}{T}}$. Then for all $T \geq \log(K)$, the regret of Hedge satisfies,

$$R_T(\pi^{\text{Hedge}}, \ell) \leq 2\sqrt{T \log(K)}.$$

Proof. Let us first check the conditions for the lemma from the previous slide,

$$T \geq \log(K) \implies \eta \leq 1, \qquad \ell_t \in [0,1]^K \implies p_t^\top \ell_t \leq 1.$$

Then, as $\ell_t^2(a) \leq 1$ for all a, we have $p_t^\top \ell_t^2 \leq 1$. Therefore, for any $p = (p_1, \dots, p_T)$ chosen by Hedge, and any action $a \in [K]$,

$$\overline{R}_T(p, \ell, a) \leq \frac{\log(K)}{\eta} + \eta T = 2\sqrt{T \log(K)}.$$

Then, $R_T(\pi^{\text{Hedge}}, \ell) = \overline{R}_T(p(\pi^{\text{Hedge}}, \ell), \ell, a^{\text{OPT}}(\ell)) \leq 2\sqrt{T \log(K)}$. □

Proof of Hedge lemma

Recall, Hedge Lemma. Let $p = (p_1, \dots, p_T)$ be the sequence of probability vectors chosen by Hedge with learning rate $\eta \in [0, 1]$. Then, for any set of loss vectors $\ell = (\ell_1, \dots, \ell_T)$, where $\ell_t \in \mathbb{R}_+^K$ and any $a \in [K]$, if $p_t^\top \ell_t \leq 1$ for all t, we have

$$\overline{R}_T(p, \ell, a) \leq \frac{\log(K)}{\eta} + \eta \sum_{t=1}^{T} p_t^\top \ell_t^2.$$

Proof. Define $\Phi_t \overset{\Delta}{=} \frac{1}{\eta} \log\left(\sum_{i=1}^K e^{-\eta L_t(i)}\right)$. Now consider,

$$\begin{aligned}
\Phi_t - \Phi_{t-1} &= \frac{1}{\eta} \log\left(\frac{\sum_{i=1}^K e^{-\eta L_t(i)}}{\sum_{i=1}^K e^{-\eta L_{t-1}(i)}}\right) \\
&= \frac{1}{\eta} \log\left(\frac{\sum_{i=1}^K e^{-\eta L_{t-1}(i)} \cdot e^{-\eta \ell_t(i)}}{\sum_{j=1}^K e^{-\eta L_{t-1}(j)}}\right) \qquad \texttt{As } L_t(i) = \sum_{s=1}^{t} \ell_s(i). \\
&= \frac{1}{\eta} \log\left(\sum_{i=1}^K p_t(i) e^{-\eta \ell_t(i)}\right) \qquad \texttt{As } p_t(i) = \frac{e^{-\eta L_{t-1}(i)}}{\sum_{j=1}^K e^{-\eta L_{t-1}(j)}}.
\end{aligned}$$

Proof of Hedge lemma (cont'd)

$$\begin{aligned}
\Phi_t - \Phi_{t-1} &= \frac{1}{\eta}\log\left(\sum_{i=1}^{K} p_t(i) e^{-\eta \ell_t(i)}\right) \\
&\le \frac{1}{\eta}\log\left(\sum_{i=1}^{K} p_t(i)\left(1-\eta\ell_t(i)+\eta^2\ell_t^2(i)\right)\right) && \texttt{See (i) below.} \\
&= \frac{1}{\eta}\log\left(1-\eta p_t^\top \ell_t + \eta^2 p_t^\top \ell_t^2\right) \\
&\le -p_t^\top \ell_t + \eta p_t^\top \ell_t^2 && \texttt{See (ii) below.}
\end{aligned}$$

(i) Using $e^{-y} \le 1 - y + y^2$ for $y \ge -1$. Applied with $y = \eta\ell_t(i) \ge 0$.

(ii) Using $\log(1+y) \le y$ for all $y > -1$. Applied with

$$y = \eta^2 p_t^\top \ell_t^2 - \eta p_t^\top \ell_t > -\underbrace{\eta}_{\le 1}\underbrace{p_t^\top \ell_t}_{\le 1} \ge -1.$$

Proof of Hedge lemma (cont'd)

$$L_T(i) = \sum_{t=1}^{T} \ell_t(i), \qquad \Phi_t \triangleq \frac{1}{\eta} \log \left(\sum_{i=1}^{K} e^{-\eta L_t(i)} \right), \qquad \Phi_t - \Phi_{t-1} \le -p_t^\top \ell_t + \eta p_t^\top \ell_t^2$$

Summing from $t = 1, \dots, T$, we have

$$\Phi_T - \Phi_0 \le -\sum_{t=1}^{T} p_t^\top \ell_t + \eta \sum_{t=1}^{T} p_t^\top \ell_t^2 .$$

Now note that, for any $a \in [K]$,

$$\Phi_0 = \frac{1}{\eta} \log \left(\sum_{i=1}^{K} e^{-\eta L_0(i)} \right) = \frac{\log(K)}{\eta} \quad \text{as } L_0 = \mathbf{0}.$$

$$\Phi_T = \frac{1}{\eta} \log \left(\sum_{i=1}^{K} e^{-\eta L_T(i)} \right) \ge \frac{1}{\eta} \log \left(e^{-\eta L_T(a)} \right) = -L_T(a) = -\sum_{t=1}^{T} \ell_t(a).$$

Proof of Hedge lemma (cont'd)

Recall, Hedge Lemma. *We need to show:* Let $p = (p_1, \ldots, p_T)$ be the sequence of probability vectors chosen by Hedge with learning rate $\eta \in [0, 1]$. Then, for any set of loss vectors $\ell = (\ell_1, \ldots, \ell_T)$, where $\ell_t \in \mathbb{R}_+^K$ and any $a \in [K]$, if $p_t^\top \ell_t \leq 1$ for all t, we have

$$\overline{R}_T(p, \ell, a) = \sum_{t=1}^{T} p_t^\top \ell_t - \sum_{t=1}^{T} \ell_t(a) \leq \frac{\log(K)}{\eta} + \eta \sum_{t=1}^{T} p_t^\top \ell_t^2.$$

This gives us,

$$-\sum_{t=1}^{T} \ell_t(a) - \frac{\log(K)}{\eta} \leq \Phi_T - \Phi_0 \leq -\sum_{t=1}^{T} p_t^\top \ell_t + \eta \sum_{t=1}^{T} p_t^\top \ell_t^2.$$

Therefore,

$$\overline{R}_T(p, \ell, a) = \sum_{t=1}^{T} p_t^\top \ell_t - \sum_{t=1}^{T} \ell_t(a) \leq \frac{\log(K)}{\eta} + \eta \sum_{t=1}^{T} p_t^\top \ell_t^2$$

□

Ch 4.3: Adversarial bandits

The adversarial bandit problem is a variant of the experts problem where the learner only observes the loss for the action she took (called bandit feedback).

- There are a set of K actions, denoted $[K]$.
- On round t, the learner chooses a distribution $p_t \in \Delta_K$. An action A_t is sampled from p_t.
- An *adversary* (environment) simultaneously picks a loss vector $\ell_t \in [0,1]^K$, where $\ell_t(i)$ is the loss for action i.
- The learner incurs loss $\ell_t(A_t)$.
- The learner observes **only** $\ell_t(A_t)$.

Regret. The regret of a policy π under losses ℓ is defined similar to the experts problem, but has an expectation around the first term (Why?):

$$R_T(\pi, \ell) = \mathbb{E}\left[\sum_{t=1}^{T} p_t^\top \ell_t\right] - \min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a)$$

The EXP3 algorithm

Key idea: We will build on Hedge, but estimate the loss vector on round t by only observing $\ell_t(A_t)$.

We will use the following estimate:

$$\widehat{\ell}_t(a) = \frac{\ell_t(a)}{p_t(a)}\mathbb{1}(A_t = a) = \begin{cases} \frac{\ell_t(a)}{p_t(a)} & \text{if } a = A_t, \\ 0 & \text{otherwise} \end{cases}$$

Here, $p_t(a)$ is the probability of choosing action a on round t in Hedge.

We will show that $\widehat{\ell}_t$ is an unbiased estimator of ℓ_t, i.e $\mathbb{E}[\widehat{\ell}_t|p_t] = \ell_t$. Intuitively, since we only see $\ell_t(a)$ when action a is picked, inflating the observed value by dividing by $p_t(a)$ ensures that the estimate is unbiased.

EXP3 (exponential weights for exploration and exploitation)

Algorithm: EXP3

```
▶ Given: time horizon $T$, learning rate $\eta$.        # Will specify $\eta$ later.
▶ Let $L_0 = \mathbf{0}_K$.        # $L_t \in \mathbb{R}_+^K$ will maintain cumulative losses.
▶ for $t = 1, \dots, T$,
    ▶ Construct $p_t \in \Delta([K])$ as follows,
          $p_t(a) \leftarrow \frac{e^{-\eta L_{t-1}(a)}}{\sum_{j=1}^{K} e^{-\eta L_{t-1}(j)}}$
    ▶ Sample $A_t \sim p_t$ and execute $A_t$.        # $\pi(\cdot|\texttt{history}) = p_t(\cdot)$.
    ▶ Observe $\ell_t(A_t)$. Update, $L_t(a) \leftarrow L_{t-1}(a) + \widehat{\ell}_t(a)$ for all $a \in [K]$. That is,
          $L_t(A_t) \leftarrow L_{t-1}(A_t) + \frac{\ell_t(A_t)}{p_t(A_t)}$,        $L_t(a) \leftarrow L_{t-1}(a)$ for all $a \neq A_t$.
```

The only difference with Hedge is in the loss update, where we use $\widehat{\ell}_t$ instead of ℓ_t.

Analysis of EXP3

Lemma. Let $\widehat{\ell}_t(a) = \frac{\ell_t(a)}{p_t(a)} \mathbb{1}(A_t = a)$ and $\widehat{\ell}_t^2 \in \mathbb{R}_+^K$ be such that $\widehat{\ell}_t^2(a) = (\widehat{\ell}_t(a))^2$. Then, for all $a \in [K]$,

1. $\mathbb{E}[\widehat{\ell}_t(a) | p_t] = \ell_t(a)$.
2. $\mathbb{E}[\widehat{\ell}_t^2(a) | p_t] = \frac{\ell_t^2(a)}{p_t(a)}$.

Proof. 1. For any $a \in [K]$, $\widehat{\ell}_t(a)$ will be $\frac{\ell_t(a)}{p_t(a)}$ with probability $p_t(a)$ (when action a is sampled, *i.e.*, $A_t = a$) and 0 with probability $1 - p_t(a)$ (when any other action is sampled).

$$\mathbb{E}[\widehat{\ell}_t(a) | p_t] = p_t(a) \cdot \frac{\ell_t(a)}{p_t(a)} + (1 - p_t(a)) \cdot 0 = \ell_t(a).$$

2. Similarly, for any $a \in [K]$,

$$\mathbb{E}[\widehat{\ell}_t^2(a) | p_t] = p_t(a) \cdot \frac{\ell_t^2(a)}{p_t^2(a)} + (1 - p_t(a)) \cdot 0 = \frac{\ell_t^2(a)}{p_t(a)}.$$

Regret bound for EXP3

$$R_T(\pi, \ell) = \mathbb{E}\left[\sum_{t=1}^{T} p_t^\top \ell_t\right] - \min_{i \in [K]} \sum_{t=1}^{T} \ell_t(i).$$

Theorem. Suppose $\ell_t \in [0,1]^K$ for all t, and we choose $\eta = \sqrt{\frac{\log(K)}{KT}}$. Then for all T, the regret of EXP3 satisfies,

$$R_T(\pi^{\mathrm{EXP3}}, \ell) \leq 2\sqrt{KT \log(K)}.$$

Remark. The regret of Hedge is $\mathcal{O}(\sqrt{T \log(K)})$, whereas for EXP3, it is $\mathcal{O}(\sqrt{KT \log(K)})$. The additional $\sqrt{K}$ factor is due to reduced (limited) feedback.

Proof of regret bound for EXP3

Recall, Hedge Lemma. Let $\lambda = (\lambda_1, \ldots, \lambda_T) \in \mathbb{R}_+^K$ be a sequence of losses. Let $p = (p_1, \ldots, p_T)$ be the sequence of probability vectors chosen by Hedge with learning rate $\eta \in [0, 1]$. For any $a \in [K]$, if $p_t^\top \lambda_t \le 1$ for all t, we have

$$\overline{R}_T(p, \lambda, a) \triangleq \sum_{t=1}^{T} p_t^\top \lambda_t - \sum_{t=1}^{T} \lambda_t(a) \le \frac{\log(K)}{\eta} + \eta \sum_{t=1}^{T} p_t^\top \lambda_t^2.$$

Proof. Let $a^{\mathrm{OPT}} = \mathrm{argmin}_{a \in [K]} \sum_{t=1}^{T} \ell_t(a)$ be the best fixed arm in hindsight.

We will apply the above lemma with $\lambda_t \leftarrow \widehat{\ell}_t$ and $a \leftarrow a^{\mathrm{OPT}}$.

Let us first verify the conditions,

$$\eta = \sqrt{\frac{\log(K)}{KT}} \le 1 \quad \text{as } K \ge 2 \text{ and } T \ge 1.$$

$$p_t^\top \widehat{\ell}_t = \sum_{a=1}^{K} p_t(a) \cdot \frac{\ell_t(a)}{p_t(a)} \mathbb{1}(A_t = a) = p_t(A_t) \cdot \frac{\ell_t(A_t)}{p_t(A_t)} = \ell_t(A_t) \le 1.$$

Proof of regret bound for EXP3 (cont'd)

Recall $\forall\, a$, *(i)* $\mathbb{E}[\widehat{\ell}_t(a)|p_t] = \ell_t(a)$, *(ii)* $\mathbb{E}[\widehat{\ell}_t^{\,2}(a)|p_t] = \frac{\ell_t^2(a)}{p_t(a)}$.

Therefore we have, by the Hedge lemma,

$$\sum_{t=1}^{T} p_t^\top \widehat{\ell}_t - \sum_{t=1}^{T} \widehat{\ell}_t(a^{\mathrm{OPT}}) \leq \frac{\log(K)}{\eta} + \eta \sum_{t=1}^{T} p_t^\top \widehat{\ell}_t^{\,2}.$$

This is a good preiminary bound, but depends on the *random* quantities $\widehat{\ell}_t$. To deal with this, let us take expectations on both sides,

$$\mathbb{E}[\mathrm{LHS}] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}[p_t^\top \widehat{\ell}_t | p_t] - \sum_{t=1}^{T} \underbrace{\mathbb{E}[\widehat{\ell}_t(a^{\mathrm{OPT}})|p_t]}_{=\ell_t(a^{\mathrm{OPT}}) \text{ by } (i)}\right]$$

Further, by *(i)* again

$$\mathbb{E}[p_t^\top \widehat{\ell}_t | p_t] = \mathbb{E}\left[\sum_{a=1}^{K} p_t(a)\widehat{\ell}_t(a) \middle| p_t\right] = \sum_{a=1}^{K} p_t(a) \underbrace{\mathbb{E}\left[\widehat{\ell}_t(a)|p_t\right]}_{=\ell_t(a) \text{ by } (i)} = p_t^\top \ell_t$$

Proof of regret bound for EXP3 (cont'd)

$$\text{Recall } \forall\, a, \qquad \textit{(i)}\ \mathbb{E}[\widehat{\ell}_t(a)|p_t] = \ell_t(a), \qquad \textit{(ii)}\ \mathbb{E}[\widehat{\ell}_t^{\,2}(a)|p_t] = \frac{\ell_t^2(a)}{p_t(a)}.$$

Therefore,

$$\mathbb{E}[\mathrm{LHS}] = \mathbb{E}\left[\sum_{t=1}^{T} p_t^\top \ell_t - \sum_{t=1}^{T} \ell_t(a^{\mathrm{OPT}})\right] = \mathbb{E}\left[\sum_{t=1}^{T} p_t^\top \ell_t\right] - \sum_{t=1}^{T} \ell_t(a^{\mathrm{OPT}}) = R_T(\pi, \ell).$$

Now consider the RHS,

$$\mathbb{E}[\mathrm{RHS}] = \frac{\log(K)}{\eta} + \eta\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}[p_t^\top \widehat{\ell}_t^{\,2} | p_t]\right].$$

By *(ii)* and as losses are bounded in $[0, 1]$,

$$\mathbb{E}[p_t^\top \widehat{\ell}_t^{\,2} | p_t] = p_t^\top \mathbb{E}[\widehat{\ell}_t^{\,2} | p_t] = \sum_{a=1}^{K} p_t(a) \cdot \frac{\ell_t^2(a)}{p_t(a)} = \sum_{a=1}^{K} \ell_t^2(a) \le K.$$

Proof of regret bound for EXP3 (cont'd)

Therefore,

$$\mathbb{E}[\mathrm{RHS}] \leq \frac{\log(K)}{\eta} + \eta K T.$$

Hence,

$$\begin{aligned} R_T(\pi, \ell) &\leq \frac{\log(K)}{\eta} + \eta K T \\ &\leq 2\sqrt{KT \log(K)} \quad \text{as } \eta = \sqrt{\frac{\log(K)}{KT}}. \end{aligned}$$

□