CS861: Theoretical Foundations of Machine Learning

Chapter 5: Online learning: Learning with expert advice & Adversarial Bandits

Kirthevasan Kandasamy
UW-Madison

Outline

1. Learning from experts
2. Adversarial bandits
3. Lower bounds for adversarial bandits
4. Contextual bandits

Ch 5.1: Learning with expert advice

Example 1. *Spam prediction.* Given a hypothesis class $\mathcal{H}$, consider the following game over T rounds:

- A learner receives an email x_t on round t.
- The learner chooses some $h_t \in \mathcal{H}$ and predicts $h_t(x_t) \in \{\texttt{spam}, \texttt{not-spam}\}$.
- The learner then sees the true label y_t and incurs loss $\mathbb{1}(h_t(x_t) \neq y_t)$.

Note that the learner can compute the loss for all $h \in \mathcal{H}$ after observing the label.

Example 2. *Weather forecasting.* Given a set of models $\mathcal{H}$,

- The learner chooses some $h_t \in \mathcal{H}$ and outputs a prediction $\widehat{y}_t$.
- The learner then observes the true weather y_t and incurs loss $\ell(\widehat{y}_t, y_t)$.

Learner knows the loss incurred by all models $h \in \mathcal{H}$ after observing the weather.

Learning with expert advice, a.k.a the "Experts problem"

Problem set up:

- There are a set of K experts, denoted $[K]$.

- On round t, the learner chooses an expert (action) $A_t \in [K]$.

- An *adversary* (environment) simultaneously, i.e without knowledge of A_t, picks a loss vector $\ell_t \in [0,1]^K$, where $\ell_t(i)$ is the loss for expert i.

- The learner incurs loss $\ell_t(A_t)$.

- The learner observes the entire loss vector ℓ_t, i.e the losses for *all* experts.

The learner observes the losses for *all* actions. This is known as *full information*.
C.f. In *bandit feedback*, we observe the losses only for the action A_t we took.

Learner's policy

A learner is characterized by a *policy* $\pi = \{\pi_t\}_{t \in \mathbb{N}}$.

Here, π_t maps the history $\{(A_s, \ell_s)\}_{s=1}^{t-1}$ to a distribution over $[K]$, and then an action is sampled from this distribution.

$$\pi_t : \left\{[K] \times [0,1]^K\right\}^{t-1} \to \Delta([K]),$$
$$\text{where, } \Delta([K]) = \{p \in [0,1]^K;\ p^\top \mathbf{1}_K = 1\}$$

The algorithm(s) we will study will be randomized (unlike UCB). In fact, deterministic policies will fail spectacularly in the adversarial setting.

Regret in the adversarial setting

Question: How do you define regret in the adversarial (nonstochastic) setting?

- Recall, in the stochastic bandit setting, we competed with $a_\star = \operatorname{argmin}_a \mathbb{E}_{X \sim \nu_i}[X]$ (we can define it similarly for stochastic full information settings).
- But in the adversarial setting, the losses could be arbitrary.

Ans: We will compete against the *best fixed action in hindsight*.

Regret in the adversarial setting (cont'd)

We will compete against the *best fixed action in hindsight*.

For a sequence of losses $\ell = (\ell_1, \ldots, \ell_T) \in [0,1]^{K \times T}$ and a sequence of actions $A = (A_1, \ldots, A_T) \in [K]^T$, define the regret as,

$$R'_T(A, \ell) = \sum_{t=1}^{T} \ell_t(A_t) \; - \; \min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a).$$

For a randomized policy π, we will define the regret as

$$R_T(\pi, \ell) = \mathbb{E}\left[R'_T(A, \ell)\right] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(A_t)\right] \; - \; \min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a).$$

where $\mathbb{E}$ is with respect to the randomness of the policy, i.e $A_t \sim \pi(\cdot | A_1, \ell_1, \ldots, A_{t-1}, \ell_{t-1})$.

Regret in the adversarial setting

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

For now, let us assume that the adversary chooses the entire loss sequence $\ell = (\ell_1, \ldots, \ell_T)$ ahead of time, possibly with knowledge of the learner's policy π. This is called an *oblivious adversary* (We will revisit this assumption later).

We wish to achieve small regret $R_T(\pi, \ell)$ for *all* loss sequences $\ell \in [0,1]^{K \times T}$.
That is, we are interested in designing π to minimize $\sup_{\ell \in [0,1]^K} R_T(\pi, \ell)$.

Designing a policy for the experts problem

- On round t, the learner chooses an expert (action) $A_t \in [K]$.
- The learner incurs loss $\ell_t(At)$, but observes the entire loss vector ℓ_t, i.e losses for *all* experts.

Most straightforward idea: Follow the leader (FTL):

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

E.g. For binary classification, FTL is simply ERM as we will choose $h_t = \operatorname{argmin}_{h \in \mathcal{H}} \sum_{s=1}^{t-1} \mathbb{1}(h(X_t) \neq Y_t)$.

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 $T/2 - 1$.

Try at home: Extend this example to show that *any* deterministic policy will fail.
Recall, we wish to bound $\sup_\ell R_T(\pi, \ell)$, so we only need to find one bad loss vector.

The Hedge algorithm

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

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

- Given: time horizon T, learning rate η. `# Will specify` η `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,

$$p_t(a) \leftarrow \frac{e^{-\eta L_{t-1}(a)}}{\sum_{j=1}^{K} e^{-\eta L_{t-1}(j)}}$$

  - Sample $A_t \sim p_t$ and execute A_t. `#` $\pi(\cdot|\texttt{history}) = p_t(\cdot)$.
  - Observe ℓ_t. Update $L_t(a) \leftarrow L_{t-1}(a) + \ell_t(a)$ for all $a \in [K]$.

Regret analysis for the experts problem: preparation

Let $p_t(\cdot) = \pi(\cdot|\{A_s, \ell_s\}_{s=1}^{t-1})$ be the probability distribution over $[K]$ from which action A_t is sampled. Then,

$$\mathbb{E}[\ell_t(A_t)|p_t] = p_t^\top \ell_t.$$

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

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

Regret analysis for the experts problem: preparation (cont'd)

$$\mathbb{E}[\ell_t(A_t)|p_t] = p_t^\top \ell_t, \qquad \overline{R}_T(p, \ell, a) \triangleq \sum_{t=1}^{T} p_t^\top \ell_t - \sum_{t=1}^{T} \ell_t(a),$$

We now have, for any policy π,

$$\begin{aligned} R_T(\pi, \ell) &= \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(A_t)\right] - \min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a) \\ &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}[\ell_t(A_t)|p_t]\right] - \sum_{t=1}^{T} \ell_t(a_\star(\ell)) \qquad \texttt{where, } a_\star(\ell) = \operatorname*{argmin}_{a \in [K]} \sum_t \ell_t(a) \\ &= \mathbb{E}\left[\overline{R}_T(p, \ell, a_\star(\ell))\right] \end{aligned}$$

Hence, if we can bound $\overline{R}_T(p, \ell, a)$ for any action $a \in [K]$, and any p chosen by π, we can bound $R_T(\pi, \ell)$.

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.

We will first bound the regret of Hedge using this lemma, and then prove this lemma.

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,

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

Proof. Let us first check the conditions,

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

Then, as $\ell_t^2(a) \leq 1$ for all a, we have $p_t^\top \ell_t^2 \leq 1$. Therefore, for any $p = (p_1, \ldots, p_T)$ chosen by Hedge,

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

Then, $R_T(\pi^{\mathrm{Hedge}}, \ell) = \mathbb{E}[\overline{R}_T(p, \ell, a_\star)] \leq 2\sqrt{T \log(K)}$. □

Proof of Hedge lemma

Recall, Hedge 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

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

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

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

Proof of Hedge lemma (cont'd)

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

(i) Using $e^{-y} \leq 1 - y + y^2$ for $y \geq -1$. Applied with $y = \eta \ell_t(i) \geq 0$.
(ii) Using $\log(1 + y) \leq y$ for all $y > -1$. Applied with

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

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 $t = 1, \ldots, T$, 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)

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,

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

□

Oblivious vs adaptive adversaries

We designed a policy π to minimize $\sup_\ell R_T(\pi, \ell)$ where,

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

We may have two types of adversaries:

1. Oblivious adversary: Adversary chooses the entire loss sequence $\ell = (\ell_1, \ldots, \ell_T)$ ahead of time, possibly with knowledge of the learner's policy. That is, $\ell_t(\cdot)$ is only a function of the action, i.e $\ell_t(i)$.
2. Adaptive adversary: Adversary chooses the loss on round t, after having witnessed the history $A_1, \ell_1, \ldots, A_{t-1}, \ell_{t-1}$. That is, the adversary chooses loss *functions* $\ell_t(\cdot; A_1, \ell_1, \ldots, A_{t-1}, \ell_{t-1})$.

Question: If we have an adaptive instead of an oblivious adversary:

1. Does our proof of Hedge still carry through?
2. How do you interpret the regret defined above?

Ch 3.2: 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 an action $A_t \in [K]$.
- 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. (Defined exactly as for the experts problem) For a randomized policy π, define the regret as

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

where $\mathbb{E}$ is w.r.t the randomness of π.

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 $\mathbb{E}[\widehat{\ell}_t | p_t] = \ell_t$.

EXP3 (exponential weights for exploration and exploitation)

Algorithm: EXP3

- Given: time horizon T, learning rate η. `# Will specify` η `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,

$$p_t(a) \leftarrow \frac{e^{-\eta L_{t-1}(a)}}{\sum_{j=1}^{K} e^{-\eta L_{t-1}(j)}}$$

  - 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,

$$L_t(A_t) \leftarrow L_{t-1}(A_t) + \frac{\ell_t(A_t)}{p_t(A_t)}, \qquad L_t(a) \leftarrow L_{t-1}(a) \text{ for all } a \neq A_t.$$

Question: How does EXP3 manage the exploration–exploitation trade-off?

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 $a \in [K]$

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

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

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

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 $p_t^\top \lambda_t \leq 1$ for all t, we have

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

Proof. Let $a_\star = \operatorname{argmin}_{a \in [K]} \sum_{t=1}^{T} \ell_t(a)$ be the best fixed arm in hindsight.
We will apply the above lemma with $\lambda_t \leftarrow \widehat{\ell}_t$ and $a \leftarrow a_\star$.
Let us first verify the conditions,

$$\eta = \sqrt{\frac{\log(K)}{KT}} \leq 1 \quad \texttt{as } K \geq 2 \texttt{ and } T \geq 1.$$

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

Proof of regret bound for EXP3 (cont'd)

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

Therefore we have,

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

Now, let us take expectations on both sides,

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

Further by *(i)* again

$$\mathbb{E}[p_t^\top \widehat{\ell}_t|p_t] = p_t^\top \mathbb{E}[\widehat{\ell}_t|p_t] = p_t^\top \ell_t = \mathbb{E}[\ell_t(A_t)|p_t]$$

Proof of regret bound for EXP3 (cont'd)

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

Therefore,

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

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 $[0, 1]$,

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

Proof of regret bound for EXP3 (cont'd)

Therefore,

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

Hence,

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

□

Full information vs bandit feedback

- Full information feedback: Learner observes the losses for *all* actions.
- Bandit feedback: learner observes the loss only for the action A_t she took.

	Full information feedback	bandit feedback
Stochastic	?	Stochastic bandits (UCB)
Adversarial	Experts problem (Hedge)	Adversarial bandits (EXP3)

Ch 3.3: Lower bounds for adversarial bandits

Theorem. For the adversarial multi-armed bandit problem, the minimax regret satisfies

$$\inf_{\pi} \sup_{\ell \in [0,1]^{K \times T}} R_T(\pi, \ell) \in \Omega(\sqrt{KT}).$$

- Recall the lower bound for sub-Gaussian stochastic multi-armed bandits, $\inf_\pi \sup_\nu R_T(\pi, \nu) \in \Omega(\sqrt{KT})$.
- The regret definitions are different (best action in hindsight vs expectation), but the same rate!

Lower bound for adversarial bandits (cont'd)

Proof strategy. Let π be given. Our strategy will be as follows:

1. Consider a distribution over losses in $[0,1]^{K\times T}$ and show that the expected regret of π over this distribution is large.
2. Then, there should be at least one sequence of losses (drawn from this distribution) which should have large regret.

Proof. Consider two *stochastic bandit models* $\nu^{(1)}, \nu^{(2)}$ where $\nu^{(j)} = (\nu_1^{(j)}, \dots, \nu_K^{(j)})$ and each $\nu_i^{(j)}$ has Bernoulli losses. We will choose $\nu^{(1)}, \nu^{(2)}$ based on π.

Let $\mathbb{P}^{(1)}, \mathbb{P}^{(2)}$ denote the probability law of the action-loss sequence $A_1, \ell_1(A_1), \dots, A_t, \ell_t(A_t), \dots, A_T, \ell_T(A_T)$ due to π's interaction with $\nu^{(1)}, \nu^{(2)}$ respectively.

Let $\mathbb{E}^{(1)}, \mathbb{E}^{(2)}$ denote the corresponding expectations.

Lower bound for adversarial bandits: lower bounding adversarial regret with stochastic regret

Let $\mathbb{E}_\pi$ denote the expectation with respect to the randomness in the policy. For a fixed sequence of actions $A = (A_1, \ldots, A_T)$ and losses $\ell = (\ell_1, \ldots, \ell_T)$, define

$$R'_T(A, \ell) \triangleq \sum_{t=1}^{T} \ell_t(A_t) - \min_{a \in [K]} \sum_{t=1}^{T} \ell_t(a),$$

so that $R_T(\pi, \ell) = \mathbb{E}_\pi[R'_T(A, \ell)]$.

We can now lower bound the worst case regret for π as follows,

$$\sup_{\ell \in [0,1]^{K \times T}} R_T(\pi, \ell) = \sup_{\ell \in [0,1]^{K \times T}} \mathbb{E}_\pi \left[R'_T(A, \ell) \right] \quad \geq \mathbb{E}_{j \sim \mathrm{Unif}\,(\{1,2\})} \mathbb{E}_{\ell \sim \nu^{(j)}} \mathbb{E}_\pi \left[R'_T(A, \ell) \right]$$

The last step uses max $\geq$ avg, noting that $j \sim \mathrm{Unif}\,(\{1, 2\})$ and then $\ell \sim \nu^{(j)}$ defines a distribution over $[0, 1]^{K \times T}$.

Lower bound for adversarial bandits: lower bounding adversarial regret with stochastic regret (cont'd)

Continuing

$$\sup_{\ell \in [0,1]^{K \times T}} R_T(\pi, \ell) \geq \mathbb{E}_{j \sim \mathrm{Unif}(\{1,2\})} \mathbb{E}_{\ell \sim \nu^{(j)}} \mathbb{E}_\pi \left[R'_T(\pi, \ell) \right]$$

$$= \frac{1}{2} \mathbb{E}_\pi \left[\mathbb{E}_{\ell \sim \nu^{(1)}} \left[R'_T(A, \ell) \right] \right] + \frac{1}{2} \mathbb{E}_\pi \left[\mathbb{E}_{\ell \sim \nu^{(2)}} \left[R'_T(A, \ell) \right] \right].$$

Let $\mu_j^\star = \min_{a \in [K]} \mathbb{E}_{X \sim \nu_i^{(j)}}[X]$ denote the minimum mean value in $\nu^{(j)}$. By Jensen's inequality, and the fact that the pointwise minimum is concave, we have $\mathbb{E}[\min_i z_i] \leq \min_i \mathbb{E}[z_i]$. Using this we can write,

$$\mathbb{E}_{\ell \sim \nu^{(j)}} \left[R'_T(A, \ell) \right] = \mathbb{E}_{\ell \sim \nu^{(j)}} \left[\sum_{t=1}^T \ell_t(A_t) - \min_{a \in [K]} \sum_{t=1}^T \ell_t(a) \right]$$

$$\geq \mathbb{E}_{\ell \sim \nu^{(j)}} \left[\sum_{t=1}^T \ell_t(A_t) \right] - \min_{a \in [K]} \mathbb{E}_{\ell \sim \nu^{(j)}} \left[\sum_{t=1}^T \ell_t(a) \right] = \mathbb{E}_{\ell \sim \nu^{(j)}} \left[\sum_{t=1}^T \ell_t(A_t) \right] - T \mu_j^\star.$$

Lower bound for adversarial bandits: lower bounding adversarial regret with stochastic regret (cont'd)

$$\textit{(i)} \sup_{\ell \in [0,1]^{K \times T}} R_T(\pi, \ell) \geq \frac{1}{2}\mathbb{E}_\pi\left[\mathbb{E}_{\ell \sim \nu^{(1)}}\left[R'_T(A, \ell)\right]\right] + \frac{1}{2}\mathbb{E}_\pi\left[\mathbb{E}_{\ell \sim \nu^{(2)}}\left[R'_T(A, \ell)\right]\right].$$

$$\textit{(ii)}\ \mathbb{E}_{\ell \sim \nu^{(j)}}\left[R'_T(A, \ell)\right] \geq \mathbb{E}_{\ell \sim \nu^{(j)}}\left[\sum_{t=1}^{T} \ell_t(A_t)\right] - T\mu_j^\star.$$

Now, taking $\mathbb{E}_\pi$ on both sides of *(ii)*,

$$\mathbb{E}_\pi\left[\mathbb{E}_{\ell \sim \nu^{(j)}}\left[R'_T(A, \ell)\right]\right] \overset{(a)}{\geq} \mathbb{E}^{(j)}\left[\sum_{t=1}^{T} \ell(A_t)\right] - T\mu_j^\star \overset{(b)}{=} R_T^{\mathrm{stoc}}(\pi, \nu^{(j)}).$$

(a): $\mathbb{E}^{(j)}$ is expectation under π's interaction with $\nu^{(j)}$.
(b): "stochastic bandit regret" of policy π on the *stochastic bandit model* $\nu^{(j)}$.

Combining this with *(i)*, we have

$$\sup_{\ell \in [0,1]^{K \times T}} R_T(\pi, \ell) \geq \frac{1}{2} R_T^{\mathrm{stoc}}(\pi, \nu^{(1)}) + \frac{1}{2} R_T^{\mathrm{stoc}}(\pi, \nu^{(2)}).$$

Lower bound for adversarial bandits: construction

Denote $N_{a,T} = \sum_{t=1}^{T} \mathbb{1}(A_t = a)$.

Let $\nu^{(1)}$ be defined as,

$$\nu_1^{(1)} = \mathrm{Bern}\,(1/2 - \delta), \qquad \nu_i^{(1)} = \mathrm{Bern}\,(1/2) \text{ for all } i \in \{2, \ldots, K\}$$

Here, we will specify $\delta\,(\leq 1/8)$ shortly.

As $\sum_{a=1}^{K} \mathbb{E}^{(1)}[N_{a,T}] = T$, there exists some $a' \in \{2, \ldots, K\}$ such that $\mathbb{E}^{(1)}[N_{a',T}] \leq T/(K-1)$. Define $\nu^{(2)}$ so that,

$$\nu_{a'}^{(2)} = \mathrm{Bern}\,(1/2 - 2\delta), \qquad \nu_i^{(2)} = \nu_i^{(1)} \text{ for all } i \neq a'.$$

Lower bound for adversarial bandits: lower bounding stochastic regret

$$\sup_{\ell \in [0,1]^{K \times T}} R_T(\pi, \ell) \geq \frac{1}{2} \left(R_T^{\mathrm{stoc}}(\pi, \nu^{(1)}) + R_T^{\mathrm{stoc}}(\pi, \nu^{(2)}) \right).$$

From here, the proof is very similar to our lower bound proof for stochastic bandits. From our construction,

$$R_T^{\mathrm{stoc}}(\pi, \nu^{(1)}) \geq \mathbb{P}^{(1)}\left(N_{1,T} \leq T/2\right) \frac{T\delta}{2}, \qquad R_T^{\mathrm{stoc}}(\pi, \nu^{(2)}) \geq \mathbb{P}^{(2)}\left(N_{1,T} > T/2\right) \frac{T\delta}{2}.$$

Therefore,

$$\sup_{\ell \in [0,1]^{K \times T}} R_T(\pi, \ell) \geq \frac{T\delta}{4} \left(\mathbb{P}^{(1)}\left(N_{1,T} \leq T/2\right) + \mathbb{P}^{(2)}\left(N_{1,T} > T/2\right) \right)$$

Lower bound for adversarial bandits: lower bounding stochastic regret

Bretagnolle-Huber inequality. Let P_0, P_1 be any two distributions. For any event A,

$$P_0(A) + P_1(A^c) \geq \frac{1}{2} e^{-\mathrm{KL}(P_0, P_1)}.$$

Therefore,

$$\begin{aligned}\sup_{\ell \in [0,1]^{K \times T}} R_T(\pi, \ell) &\geq \frac{T\delta}{4}\left(\mathbb{P}^{(1)}\left(N_{1,T} \leq T/2\right) + \mathbb{P}^{(2)}\left(N_{1,T} > T/2\right)\right) \\ &\geq \frac{T\delta}{8} \exp\left(-\mathrm{KL}\left(\mathbb{P}^{(1)}, \mathbb{P}^{(2)}\right)\right).\end{aligned}$$

Noting that $\mathbb{E}^{(1)}[N_{a',T}] \leq T/(K-1)$ by our construction, we have

$$\begin{aligned}\mathrm{KL}\left(\mathbb{P}^{(1)}, \mathbb{P}^{(2)}\right) &= \sum_{i=1}^{K} \mathbb{E}^{(1)}[N_{a,T}] \mathrm{KL}(\nu_i^{(1)}, \nu_i^{(2)}) = \mathbb{E}^{(1)}[N_{a',T}] \mathrm{KL}(\nu_{a'}^{(1)}, \nu_{a'}^{(2)}) \\ &\leq \frac{T}{K-1} \cdot C_1 (2\delta)^2 = C_2 \frac{T\delta^2}{K-1}.\end{aligned}$$

Lower bound for adversarial bandits: lower bounding stochastic regret

Therefore,

$$\sup_{\ell \in [0,1]^{K \times T}} R_T(\pi, \ell) \geq \frac{T\delta}{8} \exp\left(C_2 \frac{T\delta^2}{K-1} \right)$$

Choosing $\delta = \sqrt{(K-1)/T}$, we have

$$\sup_{\ell \in [0,1]^{K \times T}} R_T(\pi, \ell) \geq C_3 \sqrt{T(K-1)}.$$

□

Ch 3.4: Contextual bandits

So far, we have looked at K arms (actions) and competed against the single best action in hindsight. But the best action may depend on contextual information, which may be available to the learner.

Example. Advertising (bandits): find the best ad.
Targeted advertising (contextual bandits): find the best ad for a given query/user (context).

A policy which has good regular bandit regret may have poor performance in a real-world application.

The contextual bandit problem

- There are a set of K actions, denoted $[K]$.
- At the beginning of each round t, an *adversary* picks a context $x_t \in \mathcal{X}$. The *learner observes* x_t.
- The learner then chooses an action $A_t \in [K]$.
- The adversary simultaneously (i.e without knowledge of A_t) 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* $\ell_t(A_t)$.

Defining regret for contextual bandits

One option is to compete against the best action for the given context,

$$R_T(\pi, \ell, x) = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(A_t)\right] - \min_{e \in [K]^{\mathcal{X}}} \sum_{t=1}^{T} \ell_t(e(x_t)).$$

This is challenging if the number of possible contexts is large (possibly infinite), but also unnecessary if there are relationships between contexts. (e.g querying 'frying pan' vs 'non-stick skillet' in targeted advertising).

- You may see some contexts only once (or just a few times) so impossible to do well.

Defining regret for contextual bandits (cont'd)

Instead, we will look at a set of N "experts" who map the contexts to actions, and compete against the *best expert* in hindsight.

e.g. The experts could be ML models.

If the experts are $\{e_1, \ldots, e_N\}$, where $e_j \in [K]^{\mathcal{X}}$ for all j, then write

$$R_T(\pi, \ell, x) = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(A_t)\right] - \min_{j \in [N]} \sum_{t=1}^{T} \ell_t(e_j(x_t)).$$

We wish to design π to minimize $\sup_{\ell, x} R_T(\pi, \ell, x)$.

A simple proposition

$$R_T(\pi, \ell, x) = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(A_t)\right] - \min_{j \in [N]} \sum_{t=1}^{T} \ell_t(e_j(x_t)).$$

Can we run EXP-3 by treating the experts as actions?

- Yes, as we can define a loss vector $\widetilde{\ell}_t \in [0,1]^N$, where $\widetilde{\ell}_t(j) = \ell_t(e_j(x_t))$.
- But, this will achieve regret $\sqrt{TN\log(N)}$.
- However, we are usually interested in cases where $N \gg K$.
  - N could be as large as $K^{|\mathcal{X}|}$ (if $\mathcal{X}$ is finite).
  - If the experts are neural network models, N could be covers of the NN weights.
- We wish to reduce from $\mathrm{poly}(N)$ to $\mathrm{polylog}(N)$.

The EXP4 algorithm: Build on EXP3, but use the fact that when we observe feedback, we can discount all experts who would have chosen the same action.

The EXP4 algorithm (exponential weights for exploration and exploitation with experts)

- Given: time horizon T, learning rate η. `# Will specify` η `later.`
- Let $\widetilde{L}_0 = \mathbf{0}_N$. `#` $L_t \in \mathbb{R}_+^N$ `will maintain losses for each expert.`
- for $t = 1, \ldots, T$,
  - Observe context x_t.
  - Construct $\widetilde{p}_t \in \Delta([N])$ as follows,

$$\widetilde{p}_t(i) \leftarrow \frac{e^{-\eta L_{t-1}(i)}}{\sum_{j=1}^N e^{-\eta L_{t-1}(j)}} \qquad \text{for all experts } i \in [N].$$

  - Construct $p_t \in \Delta([K])$ via, $p_t(a) \leftarrow \sum_{j=1}^N \widetilde{p}_t(j) \mathbb{1}(e_j(x_t) = a)$.
  - Sample $A_t \sim p_t$ and execute A_t. Observe $\ell_t(A_t)$.
  - Compute action losses, $\widehat{\ell}_t(a) \leftarrow \frac{\ell_t(a)}{p_t(a)} \mathbb{1}(A_t = a)$. for all $a \in [K]$.
  - Compute expert losses, $\widetilde{\ell}(j) \leftarrow \widehat{\ell}_t(e_j(x_t))$.
  - Update cumulative losses, $\widetilde{L}_t(j) \leftarrow \widetilde{L}_{t-1}(j) + \widetilde{\ell}_t(j)$ for all $j \in [N]$.

The EXP4 algorithm: some observations

- Instead of explicitly constructing p_t, we can sample an expert E_t from $\widetilde{p}_t$ and then choose $A_t = E_t(x_t)$.

- We can write the loss update as

$$\widetilde{L}_t(j) \leftarrow \widetilde{L}_{t-1}(j) + \mathbb{1}(e_j(x_t) = A_t)\frac{\ell_t(A_t)}{p_t(A_t)}.$$

  We are using the probability of choosing A_t (via p_t), and not just the probability of choosing the relevant expert E_t.

EXP4 Regret bound

$$R_T(\pi, \ell, x) = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(A_t)\right] - \min_{j \in [N]} \sum_{t=1}^{T} \ell_t(e_j(x_t)).$$

Theorem. Suppose $\ell_t \in [0,1]^K$ for all t, and we choose $\eta = \sqrt{\frac{\log(N)}{KT}}$. Then for all $T \geq \log(N)/K$, and all $\ell \in [0,1]^{K \times T}$ and $x \in \mathcal{X}^T$, the regret of EXP4 satisfies,

$$R_T(\pi^{\mathrm{EXP4}}, \ell, x) \leq 2\sqrt{KT \log(N)}.$$

We wil use the Hedge lemma to prove this result.

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

$$\overline{R}_T(\widetilde{p}, \lambda, j) \triangleq \sum_{t=1}^{T} \widetilde{p}_t^\top \lambda_t - \sum_{t=1}^{T} \lambda_t(e_j) \leq \frac{\log(N)}{\eta} + \eta \sum_{t=1}^{T} \widetilde{p}_t^\top \lambda_t^2.$$

EXP4 Regret bound proof

Proof. Let $j_\star = \operatorname{argmin}_{j\in[N]} \sum_{t=1}^{T} \ell_t(e_j(x_t))$, be the best fixed expert in hindsight. We will apply the lemma with $j \leftarrow j_\star$, and $\lambda_t \leftarrow \widetilde{\ell}_t$. Let us first verify the conditions,

$$\eta = \sqrt{\log(N)/(KT)} \leq 1 \quad \text{as } T \geq \log(N)/K\,.$$

To verify $\widetilde{p}_t^\top \widetilde{\ell}_t \leq 1$, recall that $\widetilde{\ell}(j) \leftarrow \widehat{\ell}_t(e_j(x_t)) = \frac{\ell_t(e_j(x_t))}{p_t(e_j(x_t))} \mathbb{1}(A_t = e_j(x_t))$.
Therefore,

$$\begin{aligned}\widetilde{p}_t^\top \widetilde{\ell}_t &= \sum_{j=1}^{N} \widetilde{p}_t(j) \frac{\ell_t(e_j(x_t))}{p_t(e_j(x_t))} \mathbb{1}(A_t = e_j(x_t)) = \frac{\ell_t(A_t)}{p_t(A_t)} \sum_{j=1}^{N} \widetilde{p}_t(j) \mathbb{1}(A_t = e_j(x_t)) \\ &= \frac{\ell_t(A_t)}{p_t(A_t)} \times p_t(A_t) = \ell_t(A_t) \leq 1.\end{aligned}$$

EXP4 Regret bound: Proof (cont'd)

Recall loss update: $\widetilde{\ell}(j) \leftarrow \widehat{\ell}_t(e_j(x_t)) = \dfrac{\ell_t(e_j(x_t))}{p_t(e_j(x_t))} \mathbb{1}(A_t = e_j(x_t))$.

Now consider,

$$\mathbb{E}[\widetilde{\ell}_t(j)|\widetilde{p}_t] = p_t(e_j(x_t)) \cdot \frac{\ell_t(e_j(x_t))}{p_t(e_j(x_t))} + (1 - p_t(e_j(x_t))) \cdot 0 = \ell_t(e_j(x_t)).$$

Similarly,

$$\mathbb{E}[\widetilde{\ell}_t^2(j)|\widetilde{p}_t] = p_t(e_j(x_t)) \cdot \frac{\ell_t^2(e_j(x_t))}{p_t^2(e_j(x_t))} + (1 - p_t(e_j(x_t))) \cdot 0 = \frac{\ell_t^2(e_j(x_t))}{p_t(e_j(x_t))}$$

Remark: Here, we have $p_t(e_j(x_t)) = \sum_k \widetilde{p}_t(k)\mathbb{1}(e_k(x_t) = e_j(x_t))$ in the denominator. Naively applying EXP3 we will get $\widetilde{p}_t(j) < p_t(e_j(x_t))$ in the denominator. The estimate for the loss in EXP4 has lower variance since $\mathbb{E}[\widetilde{\ell}_t^2|\widetilde{p}_t]$ is smaller.

EXP4 Regret bound: Proof (cont'd)

$$\text{Recall,} \quad (i)\ \mathbb{E}[\widetilde{\ell}_t(j)|\widetilde{p}_t] = \ell_t(e_j(x_t)) \qquad (ii)\ \mathbb{E}[\widetilde{\ell}_t^2(j)|\widetilde{p}_t] = \frac{\ell_t^2(e_j(x_t))}{p_t(e_j(x_t))}$$

Applying the Hedge lemma with $j \leftarrow j_\star$, we get

$$\sum_{t=1}^{T} \widetilde{p}_t^\top \widetilde{\ell}_t - \sum_{t=1}^{T} \widetilde{\ell}_t(j_\star) \le \frac{\log(N)}{\eta} + \eta \sum_{t=1}^{T} \widetilde{p}_t^\top \widetilde{\ell}_t^2.$$

Let us take expectations on both sides.

$$\mathbb{E}[\text{LHS}] = \mathbb{E}\Bigg[\sum_{t=1}^{T} \mathbb{E}[\widetilde{p}_t^\top \widetilde{\ell}_t|\widetilde{p}_t] - \sum_{t=1}^{T} \underbrace{\mathbb{E}[\widetilde{\ell}_t(j_\star)|\widetilde{p}_t]}_{=\ell_t(e_{j_\star}(x_t))\text{ by }(i)} \Bigg]$$

EXP4 Regret bound: Proof (cont'd)

Recall, $\quad (i)\ \mathbb{E}[\widetilde{\ell}_t(j)|\widetilde{p}_t] = \ell_t(e_j(x_t)) \qquad (ii)\ \mathbb{E}[\widetilde{\ell}_t^2(j)|\widetilde{p}_t] = \dfrac{\ell_t^2(e_j(x_t))}{p_t(e_j(x_t))}$

By *(i)* again, we have

$$\mathbb{E}[\widetilde{p}_t^\top \widetilde{\ell}_t|\widetilde{p}_t] = \widetilde{p}_t^\top \mathbb{E}[\widetilde{\ell}_t|\widetilde{p}_t] = \sum_{j=1}^N \widetilde{p}_t(j)\ell_t(e_j(x_t)) = \sum_{j=1}^N \widetilde{p}_t(j) \sum_{a=1}^K \ell_t(a)\mathbb{1}(a = e_j(x_t))$$

$$= \sum_{a=1}^K \ell_t(a) \underbrace{\sum_{j=1}^N \widetilde{p}_t(j)\mathbb{1}(a = e_j(x_t))}_{=p_t(a)} = p_t^\top \ell_t = \mathbb{E}[\ell_t(A_t)|p_t].$$

Therefore,

$$\mathbb{E}[\mathrm{LHS}] = \mathbb{E}\left[\sum_{t=1}^T \mathbb{E}[\ell_t(A_t)|\widetilde{p}_t] - \sum_{t=1}^T \ell_t(e_{j_\star}(x_t))\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^T \ell_t(A_t)\right] - \min_{j\in[N]} \sum_{t=1}^T \ell_t(e_j(x_t)) = R_T(\pi^{\mathrm{EXP4}}, \ell, x).$$

EXP4 Regret bound: Proof (cont'd)

Recall, *(i)* $\mathbb{E}[\widetilde{\ell}_t(j)|\widetilde{p}_t] = \ell_t(e_j(x_t))$ *(ii)* $\mathbb{E}[\widetilde{\ell}_t^2(j)|\widetilde{p}_t] = \dfrac{\ell_t^2(e_j(x_t))}{p_t(e_j(x_t))}$

$$\textit{(iii)}\ \sum_{t=1}^{T} \widetilde{p}_t^\top \widetilde{\ell}_t - \sum_{t=1}^{T} \widetilde{\ell}_t(e_\star) \le \frac{\log(N)}{\eta} + \eta \sum_{t=1}^{T} \widetilde{p}_t^\top \widetilde{\ell}_t^2. \quad \text{(Hedge inequality)}$$

Now consider the RHS of *(iii)*,

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

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

$$\begin{aligned}\mathbb{E}[\widetilde{p}_t^\top \widetilde{\ell}_t^2 | \widetilde{p}_t] &= \sum_{j=1}^{N} \widetilde{p}_t(j) \frac{\ell_t^2(e_j(x_t))}{p_t(e_j(x_t))} = \sum_{j=1}^{N} \widetilde{p}_t(j) \sum_{a=1}^{K} \frac{\ell_t^2(a)}{p_t(a)} \mathbb{1}(e_j(x_t) = a) \\ &= \sum_{a=1}^{K} \frac{\ell_t^2(a)}{p_t(a)} \sum_{j=1}^{N} \widetilde{p}(j) \mathbb{1}(e_j(x_t) = a) = \sum_{a=1}^{K} \frac{\ell_t^2(a)}{p_t(a)} p_t(a) \le K.\end{aligned}$$

EXP4 Regret bound: Proof (cont'd)

$$\sum_{t=1}^{T} \widetilde{p}_t^\top \widetilde{\ell}_t - \sum_{t=1}^{T} \widetilde{\ell}_t(e_\star) \le \frac{\log(N)}{\eta} + \eta \sum_{t=1}^{T} \widetilde{p}_t^\top \widetilde{\ell}_t^2. \quad \text{(Hedge inequality)}$$

Therefore,

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

We have,

$$\begin{aligned} R_T(\pi^{\mathrm{EXP4}}, \ell, x) &\le \frac{\log(N)}{\eta} + \eta K T \\ &\le 2\sqrt{KT \log(N)} \quad \text{as } \eta = \sqrt{\log(N)/(KT)}. \end{aligned}$$