Advanced Topics in Machine Learning and Algorithmic Game Theory

Fall semester, 2011/12

Lecture 4: Multiarmed Bandit in the Adversarial Model

Lecturer: Yishay Mansour *Scribe: Shai Vardi*

4.1 Lecture Overview

In this lecture we turn our attention to the *multiarmed bandit* (MAB) model. In this model, there is a set N of actions from which the player has to choose in every step $t \in T$. After choosing the action, the player can see the loss of her action, but not the losses of the other possible actions. Notice the difference from the *full information* (FI) model (the model we used in all previous lectures), in which after the player chooses the action, she sees the losses of all possible actions. MAB models many real-life situations. For example, in choosing a route to take to work, the driver knows only her loss (how long it took to get to work), but not the losses of the other actions (how long it would have taken had she taken another route).
Today we will look at three regret minimization algorithms for MAB problems:

1. In section 4.3, we look at a general reduction from any FI regret minimization algorithm to a MAB regret minimization algorithm that guarantees a regret bound of $O(T^{2/3})$.

2. In section 4.5, we look at the *EXP3* algorithm which was introduced by Auer et al.[1], which gives a regret bound of $O(T^{1/2})$.

3. In section 4.6, we look at the *EXP4* algorithm, also by Auer et al.[1], which is a modification of the EXP3 algorithm for the expert problem where there are many more experts than actions. EXP4 also guarantees a regret bound of $O(T^{1/2})$, but achieves a better asymptotic bound with respect to the number of actions.

4.2 Exploration vs. Exploitation

The main difficulty with multiarmed bandit problems (in both the stochastic and the adversarial model) is that we don't have all of the information that we have in the full information model. Specifically - without trying an action, we have no information at all about its loss. This fact alone gives us a trivial bound on the regret: given that there are N possible actions, the regret is at least $N-1$, as we have to try each action at least once. Whereas in the FI model we were able to achieve a regret of $O(\sqrt{T \log N})$, in the MAB model it is impossible to have a logarithmic dependency on N.

This difficulty can be viewed as a trade-off between *exploration* and *exploitation*. On one hand, we need to try all the different actions. On the other hand, we would like to use the information we have so far to minimize the regret.

Clearly, in order to minimize regret, we need information about the loss of each action. In the FI model, we simply see the loss of each action at every step. How can we estimate the loss of each action in the MAB model? We will look at two methods of estimating the losses:

1. Separating exploration and exploitation. We have steps which we allocate specifically to exploration, in which we do not exploit any knowledge from previous steps.

2. Combining exploration and exploitation. In this case we use *importance sampling* - we sample actions based on their importance, at the same time making sure that all actions have a non-negligible probability of being sampled.

4.2.1 Importance sampling

Given a distribution D over our actions, we choose action i with probability p_i and we see a loss l_i. We would like to use the ratio l_i/p_i to update i's loss. Sampling according to this ratio is called *importance sampling*. Notice that $\mathbb{E}_{i\sim p}[l_i/p_i] = \sum_i p_i li/p_i = \sum_i l_i$, so if, for example, we know that for all $i \neq j$, $l_i = 0$ and $l_j = 1$, this expected loss is exactly the value we wish to find.

4.3 Reduction from Full Information

We would like to take a FI algorithm and use it to obtain an algorithm to the case of partial information. Given a regret minimization FI algorithm $\mathcal{A}$, our reduction, uses the following interaction with $\mathcal{A}$: at specific times, we give $\mathcal{A}$ a full vector of losses l (that is $l = \{l(1), l(2), \cdots l(N)\}$, where $l(i)$ is the loss of action i). $\mathcal{A}$ returns a probability distribution p over the actions.

We divide the time to T/Z blocks $\mathcal{B}$ of size Z. At the end of block $\mathcal{B}_\tau$, we give $\mathcal{A}$ the losses l_τ and receive the distribution p_τ from $\mathcal{A}$ (we call p_τ $\mathcal{A}$'s *recommendation*). In block $\mathcal{B}_{\tau+1}$, we use the latest recommendation p_τ (this is the exploitation part of the algorithm), interspersed with sampling each of the N actions once (the exploration part). Specifically, in block $\mathcal{B}_{\tau+1}$, we sample once for each action, at a time selected uniformly at random. For the rest of the block ($T/Z - N$ time slots), we sample according to p_τ. At the end of the block, we give $\mathcal{A}$ the losses of all actions we sampled in our exploration phase.
Label the loss of action i in block $\mathcal{B}_\tau$ by $x_{i,\tau}$. ($x_{i,\tau}$ is the actual loss of action i from

the exploration phase). It is easy to see that the expectation of the loss is

$$\mathbb{E}[x_{i,\tau}] = \frac{1}{Z}\sum_{t\in B_\tau} l_t(i), \tag{4.1}$$

where $l_t(i)$ is the loss of action i at time t. We take $x_{i,\tau}$ as a representative for the loss of action i in block $\mathcal{B}_\tau$. The FI algorithm $\mathcal{A}$ receives a loss vector $l_\tau = \{x_{1,\tau}, x_{2,\tau}, \cdots, x_{N,\tau}\}$ at the end of each block $\mathcal{B}_\tau$, for a total of T/Z such loss vectors. Thus, the output of $\mathcal{A}$ at the end of block τ is as if $\mathcal{A}$ played a full information game with T/Z steps, and returned, at step τ a probability distribution p_τ.

What is the expected cost of our online MAB algorithm?

$$\begin{aligned}
\mathbb{E}[OnlineMAB] &= \sum_{\tau=1}^{T/Z}\sum_{t\in B_\tau}\sum_{i=1}^{N} p_t(i)l_t(i) \\
&\leq \sum_{\tau=1}^{T/Z}\sum_{i=1}^{N} p_\tau(i)L_\tau(i) + \sum_{\tau}\sum_{i} x_{i,\tau} \qquad (4.2)\\
&\leq \sum_{\tau=1}^{T/Z}\sum_{i=1}^{N} p_\tau(i)L_\tau(i) + \frac{T}{Z}\cdot N \qquad (4.3)
\end{aligned}$$

Notice the inequality (4.2) stems from the fact that we may count the loss twice for actions at the points of exploration. In inequality (4.3), the left part represents the exploitation phase, while the right part represents the exploration phase. We notice now that we made two assumptions that we have not mentioned so far:

1. The losses for each action are bounded at each time are taken from the interval $[0, 1]$. Hence we can bound the loss of the exploration phase by the number of points of exploration.

2. The adversary must choose in advance the losses for each action. We assume that the losses in each block were already decided upon at the start of the block.

The regret minimization algorithm $\mathcal{A}$ only sees the series x_τ. Since $\mathcal{A}$ is a regret-minimization algorithm for x_τ, we have

$$\sum_{\tau}\sum_{i} p_\tau(i)x_\tau(i) \leq \sum_{\tau} x_\tau(j) + R(T/Z), \tag{4.4}$$

where the left side is the loss of $\mathcal{A}$ and j is the best action. Notice that both $x_\tau(i)$ and $x_\tau(j)$ are random variables. Similarly to equation (4.1),

$$\mathbb{E}[x_\tau(j)] = \frac{1}{Z}L_\tau(j),$$

and so we calculate the expectation of both sides of equation (4.4):

$$\sum_\tau \sum_i p_\tau(i)\mathbb{E}[x_\tau(i)] \leq \mathbb{E}[\sum_\tau x_\tau(j)] + R(T/Z),$$
$$\frac{1}{Z}\sum_\tau \sum_i p_\tau(i)L_\tau(i) \leq \frac{1}{Z}\sum_\tau L_\tau(j)] + R(T/Z),$$
$$\mathbb{E}[\text{Online exploitation}] \leq \ L_j + ZR(T/Z)$$

And therefore, adding the loss of the exploration phase, we get

$$\mathbb{E}[\text{Online MAB}] \leq \ L_j + ZR(T/Z) + NT/Z$$

Our regret in this case is

$$ZR(T/Z) + NT/Z,$$

where $R(T/Z) = 2\sqrt{\frac{T}{Z}\log N}$.
We would like to choose a Z to minimize the regret. Optimizing, we get

$$Z = \frac{T^{1/3}N^{2/3}}{\log^{1/3} N}.$$

And so

$$\text{MAB Regret} = 2T^{2/3}N^{1/3}\log^{1/3} N.$$

Notice that the variable we most care about is T, and the regret is proportional to $T^{2/3}$, which is not as good as the regret bound for the FI model . This makes sense, as we make a sacrifice for exploration. Notice also that we assume that $T/Z >> N$. This is a reasonable assumption to make, because usually in regret minimization algorithms, we make the assumption that $T >> N$.

4.3.1 Many experts, few actions

What happens when we have many more experts than actions? Let's say we have N experts and A actions, where $N >> A$. For example, we have $N = 1,000$ weathermen and $A = 2$ possible actions: to take an umbrella to work or not to take an umbrella to work. We can improve the analysis of our regret bound:

In the exploration stage, we need to explore each action once, so our loss for exploration is $A\frac{T}{Z}$. The exploitation phase achieves the same bound, as the FI algorithm calculates the loss for each expert, and so the regret for the FI algorithm remains $ZR(T/Z)$. Overall, our MAB online algorithm has regret

$$A\frac{T}{Z} + ZR(\frac{T}{Z})$$

$ZR(\frac{T}{Z}) = 2\sqrt{TZ\log N}$ and so

$$Z = T^{1/3}A^{1/2}\log^{-1/3} N$$

And so the regret is

$$2T^{2/3}A^{1/2}\log^{1/3} N,$$

Which is dependent on $\log N$ instead of N.
Notice that even in this case, the dependence is still on $T^{2/3}$. When attempting to minimize regret, we usually believe that we can reach a bound of $T^{1/2}$, and so we are not completely happy with this bound. We will show how to achieve a better regret.

4.4 Importance Sampling

First, we give a short background to the use of *importance sampling*:

Assume that we can sample a random variable X from distribution D, and wish to "transfer" it to a distribution Q. We will look at a random variable Y, where

$$Y = X\frac{Q(X)}{D(X)}$$

We would like to know X's expectation over Q:

$$\begin{aligned}\mathbb{E}_{X\sim Q}[X] &= \sum Q(X)X \\ &= \sum \frac{Q(X)}{D(X)}XD(X) \\ &= \mathbb{E}_D[Y]\end{aligned}$$

This method has a lot of problems (which we will not get into in this lecture). For example, what happens if $D(X)$ is very small?

4.5 EXP3

4.5.1 The idea

- We run the *exponential weights* regret minimization algorithm.
- At time t, we choose actions according to the distribution p_t - that is, we choose action i with probability $p_t(i)$.
- We receive a profit for action i - $g_t(i)$. (Note that we will be maximizing profits rather than minimizing losses.)
- We update the estimate for the sum of profits of action i by $\frac{g_t(i)}{p_t(i)}$.

This guarantees that at all times, the expectation of the profit of i is roughly the sum:

$$\mathbb{E}[\text{estimate of } i\text{'s profit}] = \sum_t \frac{g_t(i)}{p_t(i)}$$

4.5.2 The EXP3 algorithm

The EXP3 algorithm is based on the the exponential weights regret minimization algorithm. We have a learning parameter $\eta \in [0,1]$. Each action is assigned a weight w_i, which is initialized to 1, i.e. $\forall i \in [1,2,\cdots,N], w_i(1) = 1$. We define W_t to be the sum of the weights of the actions at time t:

$$W_t = \sum_{i=1}^{N} w_i(t)$$

We use W_t to normalize the probabilities p_i of the actions.

At time t,

1. We let
$$p_i(t) = (1-\eta)\frac{w_i(t)}{W_t} + \eta\frac{1}{N}.$$
That is, $p_i(t)$ is proportional to the relative weight of i with a small correction to ensure that p_i is never too close to 0.
2. We choose an action i_t (a random variable) according to the distribution $p_1(t),\cdots,p_N(t)$.
3. We receive a profit $g_{i_t}(t) \in [0,1]$.
4. We define
$$\hat{g}_j(t) = \begin{cases} g_j(t)/p_j(t) & \text{if } j = i_t, \\ 0 & \text{otherwise} \end{cases}$$
5. We update the weights of the actions:
$$w_j(t+1) = w_j(t)e^{\eta\hat{g}_j(t)/N}$$

To summarize, we have weights that give us an estimate to how good the actions are. The actions are chosen with probability relative to their weights, and the weights are updated in an exponential fashion.

4.5.3 Bounding the regret

Before presenting our main theorem for this section, we give a few definitions.

Define $G_j = \sum_t g_j(t)$ - the total profit of action i.

Let $G_{max} = \max_j G_j$, and let $G^* \geq G_{max}$, be some value that is larger than the largest profit (we assume that G_{max} can be upper bounded).
The profit of EXP3 is

$$G_{EXP3} = \sum_t g_{i_t}(t)$$

Where $g_{i_t}(t)$ is a random variable which is dependent on all the previous choices, i.e., on all g_{i_τ} for $\tau < t$.

Theorem 4.1 *For any loss sequence, EXP3 guarantees,*

$$G_{max} - \mathbb{E}[G_{EXP3}] \leq 2\eta G_{max} + \frac{N \ln N}{\eta} \leq 2\sqrt{2G^* N \ln N}$$

The second inequality is due to optimizing η by $\eta = \sqrt{\frac{N \ln N}{2G^*}}$. We replaced G_{max} by G^* because when we calculate η, we don't know G_{max}. For $G^* \leq T$, we get that the regret is $O(\sqrt{TN \ln N})$.

To prove Theorem 4.1, we first present the following lemma, and prove the theorem assuming the lemma is true. We then prove the lemma.

Lemma 4.5.1 *In each execution of* EXP3*,*

$$G_{EXP3} \geq (1-\eta) \sum_{t=1}^{T} \hat{g}_j(t) - \frac{N \ln N}{\eta} - \frac{\eta}{N} \sum_{t=1}^{T} \sum_{i=1}^{N} \hat{g}_i(t), \tag{4.5}$$

where j is any action (and so, of course, it can be taken to be the action with the largest profit).

Proof of theorem 4.1

First we compute the expectation of $\hat{g}_i$:

$$\begin{aligned} \mathbb{E}[\hat{g}_i(t) | i_1, \cdots i_{t-1}] &= \mathbb{E}[p_i(t) \cdot \frac{g_i(t)}{p_i(t)}] \\ &= g_i(t) \end{aligned} \tag{4.6}$$

This means that the expectation of $\hat{g}_i(t)$ is independent of the history. We now take the expectation of equation (4.5):

$$\begin{aligned} \mathbb{E}[G_{EXP3}] &\geq (1-\eta)\mathbb{E}[\sum_{t=1}^{T} \hat{g}_j(t)] - \frac{N \ln N}{\eta} - \frac{\eta}{N} \sum_{t=1}^{T} \sum_{i=1}^{N} \mathbb{E}[\hat{g}_i(t)] \\ &= (1-\eta) G_j - \frac{N \ln N}{\eta} - \frac{\eta}{N} \sum_{t=1}^{T} \sum_{i=1}^{N} g_i(t) \qquad (4.7) \\ &= (1-\eta) G_j - \frac{N \ln N}{\eta} - \frac{\eta}{N} \sum_{k=1}^{N} G_k \\ &\geq (1-\eta) G_{max} - \frac{N \ln N}{\eta} - \eta G_{max} \end{aligned}$$

For (4.7), notice that $\mathbb{E}[\sum_{t=1}^{T} \hat{g}_j(t)]$ is exactly the profit of action j, i.e. G_j.

From this, we get that

$$G_{max} - \mathbb{E}[G_{EXP3}] \leq \frac{N \ln N}{\eta} + 2\eta G_{max}$$

Therefore, given Lemma 4.5.1, we get the bound on the regret of Theorem 4.1. We now turn to the proof of the lemma.

Proof of lemma 4.5.1

We use the following properties:

1. Bounding the value of $\hat{g}_i$:

$$\hat{g}_i(t) \leq \frac{g_j(t)}{p_j(t)} \leq \frac{1}{p_j(t)} \leq \frac{N}{\eta}. \tag{4.8}$$

   (Because the profit is bounded by 1.)

2. Computing the expectation of $\hat{g}_i$:

$$\sum_{i=1}^{N} p_i(t)\hat{g}_i(t) = p_{i_t}(t)\frac{g_{i_t}(t)}{p_{i_t}} = g_{i_t} \tag{4.9}$$

3. Bounding the variance of $\hat{g}_i$:

$$\begin{aligned} \sum_{i=1}^{N} p_i(t)(\hat{g}_i(t))^2 &= p_{i_t}(t)\frac{g_{i_t}}{p_{i_t}} \cdot \hat{g}_{i_t}(t) \\ &= g_{i_t}(t) \cdot \hat{g}_{i_t}(t) \\ &\leq \hat{g}_{i_t}(t) \\ &= \sum_{i=1}^{N} \hat{g}_i(t) \end{aligned} \tag{4.10}$$

   The last inequality is because the profits are bounded by 1.

Recall that $W_t = \sum_{i=1}^{N} w_i(t)$. We want to bound $\frac{W_{T+1}}{W_1}$, or specifically $\ln(\frac{W_{T+1}}{W_1})$ from above and below. Note that $W_1 = N$.

Lower bound

$$\ln \frac{W_{T+1}}{W_1} \geq \ln \frac{w_j(T+1)}{N} = \frac{\eta}{N} \sum_{t=1}^{T} \hat{g}_j(t) - \ln N$$

(From the exponential updating of the weights).

Upper bound

$$\ln \frac{W_{T+1}}{W_1} = \sum_{t=1}^{T} \ln \frac{W_{t+1}}{W_t}$$

Because all the entries in the telescopic sum cancel each other out except for the first and last. Now we look at each entry in the telescopic sum:

$$\frac{W_{t+1}}{W_t} = \sum_{i=i}^{N} \frac{w_i(t+1)}{W_t} = \sum_{i=1}^{N} \frac{w_i(t)}{W_t} e^{\frac{\eta}{N}\hat{g}_i(t)} \tag{4.11}$$

From the definition of $p_i(t)$, we get that $\frac{w_i(t)}{W_t} = \frac{p_i(t)-\eta/N}{1-\eta}$. From (4.8) we know that $\frac{\eta}{N}\hat{g}_i(t) \leq 1$, and so we can use the inequality $e^z \leq 1 + z + z^2$, which is true for $z \leq 1$. So

$$e^{\frac{\eta}{N}\hat{g}_i(t)} \leq 1 + \frac{\eta}{N}\hat{g}_i(t) + \frac{\eta^2}{N^2}\hat{g}_i^2(t) \tag{4.12}$$

From (4.11) and (4.12), we get that

$$\begin{aligned}
\frac{W_{t+1}}{W_t} &\leq \sum_{i=1}^{N} \left(\frac{p_i(t) - \eta/N}{1-\eta} \right) \cdot \left(1 + \frac{\eta}{N}\hat{g}_i(t) + \frac{\eta^2}{N^2}\hat{g}_i^2(t) \right) \\
&\leq \sum_{i=1}^{N} \frac{p_i(t) - \eta/N}{1-\eta} + \sum_{i=1}^{N} \frac{p_i(t)}{1-\eta}\frac{\eta}{N}\hat{g}_i(t) + \sum_{i=1}^{N} \frac{p_i(t)}{1-\eta}\frac{\eta^2}{N^2}\hat{g}_i^2(t) \\
&\leq 1 + \sum_{i=1}^{N} \frac{p_i(t)\eta}{N(1-\eta)}\hat{g}_i(t) + \sum_{i=1}^{N} \frac{p_i(t)\eta^2}{N^2(1-\eta)}\hat{g}_i^2(t) \\
&\leq 1 + \frac{\eta}{N(1-\eta)} g_{i_t}(t) + \frac{\eta^2}{N^2(1-\eta)} \sum_{i=1}^{N} \hat{g}_i^2(t)
\end{aligned} \tag{4.13}$$

The second inequality is due to the removal of some negative terms. The third inequality stems from the fact that $\sum_{i=1}^{N} \frac{p_i(t) - \eta/N}{1-\eta} = 1$. The last inequality is from properties 2 (equation (4.9)) and 3 (equation (4.10)).
We take the logarithm of equation (4.13), and use the inequality $\ln(1+x) \leq x$, and have,

$$\ln \frac{W_{t+1}}{W_t} \leq \frac{\eta}{N(1-\eta)} g_{i_t}(t) + \frac{\eta^2}{N^2(1-\eta)} \sum_{i=1}^{N} \hat{g}_i(t)$$

Now we sum over all time steps:

$$\sum_{t=1}^{T} \ln \frac{W_{t+1}}{W_t} \leq \frac{\eta}{N(1-\eta)} \sum_{t=1}^{T} g_{i_t}(t) + \frac{\eta^2}{N^2(1-\eta)} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{g}_i(t)$$

Noticing that $\sum_{t=1}^{T} g_{i_t}(t) = G_{EXP3}$ and combining the upper and lower bounds, we get that,

$$\frac{\eta}{N}\sum_{t=1}^{T}\hat{g}_j(t) - \ln N \leq \ln\frac{W_{T+1}}{W_1} \leq \frac{\eta}{N(1-\eta)}G_{EXP3} + \frac{\eta^2}{N^2(1-\eta)}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{g}_i(t) \quad (4.14)$$

And, multiplying both sides by $\frac{N(1-\eta)}{\eta}$, we get

$$G_{EXP3} \geq (1-\eta)\sum_{t=1}^{T}\hat{g}_j(t) - \frac{N\ln N}{\eta} - \frac{\eta}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\hat{g}_i(t)$$

Which proves Lemma 4.5.1 and thus Theorem 4.1.

We notice that the lemma revolves around how the sum of the weights changes. We can see that the ratio $\frac{W_{t+1}}{W_t}$ grows in proportion to our profit and a small quadratic term.

To summarize, we showed that EXP3 reaches a regret bound of $O(\sqrt{T})$ by combining the exploration and exploitation stages of the MAB algorithm.

4.6 EXP4

We can improve EXP3 in the case that we have many more experts than actions. We take EXP3 and modify it to *EXP4*, for the case of N experts and A actions, where $N >> A$.
In each time step, each expert i gives a distribution $\beta_i(t)$ over the actions. We give weights to the experts, and infer the weights of the actions:

The weight of expert i at time t is $w_i(t)$.
The weight of action j at time t is $\sum_{i=1}^{N} w_i(t)\beta_{i,j}(t)$.
When updating, we will use $\hat{y}_i(t)$ for the experts (instead of $\hat{g}_i(t)$ which we used in in EXP3):

$$\hat{y}_i(t) = \beta_i(t)\cdot\hat{g}(t) = \sum_{j=1}^{A}\beta_{i,j}(t)\cdot\hat{g}_j(t)$$

We now present the algorithm EXP4:

4.6.1 The EXP4 Algorithm

We initialize $\eta \in [0,1]$ and $\forall i, w_i(1) = 1$.
At time t, we

1. Receive (from the experts) the vectors β_i,
2. Calculate $W_t = \sum_{i=1}^{N} w_i(t)$,
3. Calculate, for each action j, the probability
$$p_j(t) = (1-\eta)\sum_{i=1}^{N} \frac{w_i(t)\beta_{i,j}(t)}{W_t} + \frac{\eta}{A},$$
4. Choose action j_t according to $p(t)$,
5. Receive a profit for the action $g_{j_t}(t) \in [0,1]$,
6. Calculate
$$\hat{g}_k(t) = \begin{cases} g_k(t)/p_k(t) & \text{if } k = j_t, \\ 0 & \text{otherwise} \end{cases}$$
7. For each expert $i = 1, \cdots, N$, set
$$\begin{aligned} \hat{y}_i(t) &= \beta_i(t) \cdot \hat{g}(t) \\ &= \sum_{j=1}^{A} \beta_{i,j}(t) \cdot \hat{g}_j(t) \end{aligned}$$
8. Update the weight of each expert:
$$w_i(t+1) = w_i(t) \cdot e^{\frac{\eta}{A}\hat{y}_i(t)}$$

4.6.2 Bounding the regret

Similarly to EXP3, we first state our theorem. We then state the lemma and show why it implies the theorem. Finally, we prove the lemma.

Theorem 4.2 *For any loss sequence, EXP4 guarantees,*

$$G_{max} - \mathbb{E}[G_{EXP4}] \leq 2\eta G_{max} + \frac{A\log N}{\eta} \leq 2\sqrt{2G^* A \log N}$$

Lemma 4.6.1

$$G_{EXP4} \geq (1-\eta)\sum_{t=1}^{T} \hat{y}_j(t) - \frac{A\log N}{\eta} - \frac{\eta}{A}\sum_{t-1}^{T}\sum_{i=1}^{A} \hat{g}_i(t)$$

Proof of theorem 4.2

We calculate the expectation of $\hat{y}_i(t)$:

$$\begin{aligned}\mathbb{E}[\hat{y}_i(t)] &= \mathbb{E}[\sum_{k=1}^{A} \beta_{i,k}(t)\hat{g}_k(t)] \\ &= \sum_{k=1}^{A} \beta_{i,k}(t) g_k(t) \\ &\triangleq y_i(t)\end{aligned}$$

The profit of expert i is $G_i = \sum_{t=1}^{T} y_i(t)$.

$$\frac{1}{A}\mathbb{E}[\sum_{t=1}^{T}\sum_{j=1}^{A}\hat{g}_i(t)] = \sum_{t=1}^{T}\frac{1}{A}\sum_{i=1}^{A} g_j(t) \leq \max_{1\leq i\leq N}\sum_{t=1}^{T} y_i(t) = G_{best},$$

where G_{best} is the profit of the best expert. And so

$$\begin{aligned}\mathbb{E}[G_{EXP4}] &\geq (1-\eta)\sum_{t=1}^{T} y_j(t) - \frac{A\log N}{\eta} - \eta G_{best} \\ &= (1-\eta)G_j - \frac{A\log N}{\eta} - \eta G_{best} \\ &= G_j - \frac{A\log N}{\eta} - 2\eta G_{best}\end{aligned}$$

As this is true for any expert j, the theorem follows. □

Proof of lemma 4.6.1

We set $q_i(t) = \frac{w_i(t)}{W_t}$ - the relative weight of expert i at time t. As before,

$$\sum_{t=1}^{T}\sum_{i=1}^{N} q_i(t)\hat{y}_t(t) \geq \sum_{t=1}^{T}\hat{y}_k(t) - \frac{A\ln N}{\eta} - \frac{\eta}{A}\sum_{t=1}^{T}\sum_{i=1}^{N} q_i(t)(\hat{y}_i(t))^2$$

As before, we need to bound the rightmost and leftmost terms of the above inequality:

The leftmost

$$\begin{aligned}\sum_{i=1}^{N} q_i(t)\hat{y}_i(t) &= \sum_{i=1}^{N} q_i(t)\left(\sum_{j=1}^{A}\beta_{i,j}(t)\hat{g}_j(t)\right)\\ &= \sum_{j=1}^{A}\left(\sum_{i=1}^{N} q_i(t)\beta_{i,j}(t)\right)\hat{g}_j(t)\\ &= \sum_{j=1}^{A}\frac{p_j(t)-\eta/A}{1-\eta}\hat{g}_j(t)\\ &\leq \frac{g_j(t)}{1-\eta}.\end{aligned}$$

The rightmost

$$\begin{aligned}\sum_{i=1}^{N} q_i(t)(\hat{y}_i(t))^2 &= \sum_{i=1}^{N} q_i(t)\cdot(\beta_{i,i_t}(t)\hat{g}_{i_t}(t))^2\\ &\leq (\hat{g}_{i_t}(t))^2\left(\frac{p_{i_t}(t)-\eta/A}{1-\eta}\right)\\ &\leq \frac{\hat{g}-i_t(t)}{1-\eta}\end{aligned}$$

And so:

$$\sum_{t=1}^{T} g_{i_t}(t) \geq (1-\eta)\sum_{t=1}^{T}\hat{y}_k(t) - \frac{A\ln N}{\eta} - \frac{\eta}{A}\sum_{t=1}^{T}\sum_{j=1}^{A}\hat{g}_j(t)$$

The left side of the inequality is exactly G_{EXP4}, and so the lemma proof is complete.

□

Bibliography

[1] P. Auer, N. Cesa-Bianchi, Y. Freund and R. E. Schapire, *The nonstochastic multiarmed bandit problem*, SIAM J. Comput. Vol 32, No. 1, 2002, pp. 48-77.