

## Lecture 6: November 15, 2021

Lecturer: Yishay Mansour

Scribe: Shani Jacobson, Asaf Rotenberg

## 1 Adversarial Cost: MAB

In this lecture we will see 3 algorithms for the adversarial MAB problem:

1. Reduction from MAB to full information.
2. EXP3 - algorithm for minimizing the expected regret.
3. EXP-IX - algorithm for high probability regret.

We will look at 3 different methods to construct the estimator, each with a slightly different proof.

## 2 Reminder: Randomize Weighted Majority

```

RWM( $\eta$ ),  $\eta \in [0, 1]$ 

  initialize  $\omega_1(a) = 1 \quad \forall a \in [K]$ 
  for  $t = 1, \dots, T$ :
     $W_t = \sum_a \omega_t(a)$ 
     $p_t(a) = \frac{\omega_t(a)}{W_t} \quad \forall a \in [K]$ 
    Select action  $a_t = a$  with probability  $p_t(a)$ 
    Observe  $c_t(a) \quad \forall a \in [K]$ 
    Update:
       $\omega_{t+1}(a) = (1 - \eta)^{c_t(a)} \omega_t(a) \quad \forall a \in [K]$ 
  end for

```

Figure 1: Code for RWM algorithm

**Corollary 1** Define  $L_T^{RWM} \triangleq \sum_{t=1}^T \sum_a p_t(a) c_t(a)$ , the loss of the RWM algorithm, then:

$$L_T^{RWM} \leq \begin{cases} (1 + \eta) L_T^* + \frac{\ln K}{\eta}, & \text{if } \eta \in (0, 0.5) \\ L_T^* + 2\sqrt{T \ln K}, & \text{if } \eta = \min \left\{ 0.5, \sqrt{\frac{\ln K}{T}} \right\}. \end{cases} \quad (1)$$

### 3 Reduction from MAB to full information

#### 3.1 Reduction: Concept

Given a full information algorithm with bounded regret:  $R(T, K)$ :

- Partition the time into intervals in size  $B$  (a total of  $\frac{T}{B}$  intervals).
- In each interval construct an estimator of the cost for each action:  $\hat{c}(a)$ .
- 'Send'  $\hat{c}(\cdot)$ , a vector of all the actions cost estimators, to the full information algorithm.
- 'Receive'  $p(\cdot)$ , a vector of distribution over the actions, from the full information algorithm.
- Sample 'most' of the next interval according to  $p(\cdot)$ .

Note that there is a full separation between explore and exploit.

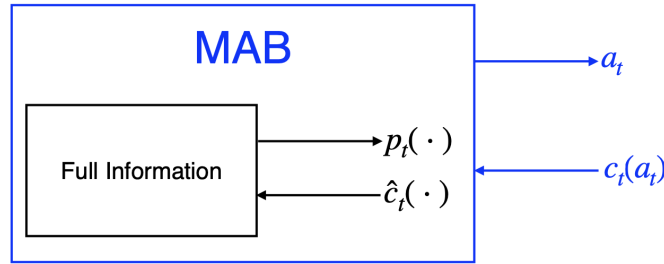


Figure 2: Reduction concept

#### 3.2 Reduction: Algorithm

1. For  $i = 1, \dots, \frac{T}{B}$ :
  - $T_i \triangleq [(i-1)B, iB]$ .
  - 'Receive'  $P_i^{FI}(\cdot)$  from the full information algorithm.
  - $\forall a$  pick time  $s_i(a)$  randomly from  $T_i$  (different time for each action).
2. Actions selection:
  - At time  $t = s_i(a)$ : play  $a_t = a$ , and get:  $c_{i,a} \triangleq c_t(a)$ .
  - Otherwise:  $a_t \sim P_i^{FI}(\cdot)$ .
3. Construction of the estimator:
  - After the end of  $T_i$ : let  $\hat{c}_i \triangleq (c_{i,a_1}, \dots, c_{i,a_k})$ .
  - 'Send'  $\hat{c}_i$  to the full information algorithm.
  - 'Receive'  $P_{i+1}^{FI}(\cdot)$  from the full information algorithm.

### 3.3 Reduction: Regret Analysis

**Theorem 2** *The expected regret of the reduction from MAB to RWM algorithm is upper bounded by  $O\left(T^{\frac{2}{3}}K^{\frac{1}{3}}\log^{\frac{1}{3}}K\right)$ .*

**Proof:** Define  $L_i(a) \triangleq \sum_{t \in T_i} c_t(a)$

$$\mathbb{E}[c_{i,a}] = \frac{1}{B} \sum_{t \in T_i} c_t(a) = \frac{1}{B} L_i(a) \quad (2)$$

This is because we selected the sampling times of the actions at random. The expected regret of the Online-MAB algorithm is upper bounded by:

$$\begin{aligned} \mathbb{E}[\text{Regret}_{\text{OnlineMAB}}] &= \sum_{i=1}^{\frac{T}{B}} \sum_{t \in T_i} \sum_a p(a_t = a) c_t(a) \\ &\leq \underbrace{\sum_{i=1}^{\frac{T}{B}} \sum_a P_i^{FI}(a) L_i(a)}_{\text{exploit}} + \underbrace{\sum_{i=1}^{\frac{T}{B}} \sum_a c_{i,a}}_{\text{explore}} \\ &\leq \sum_{i=1}^{\frac{T}{B}} \sum_a P_i^{FI}(a) L_i(a) + \frac{T}{B} K \end{aligned} \quad (3)$$

Assumptions:

- In the last transition we assumed  $c_t(a) \in [0, 1]$ .
- Oblivious adversary

In order to bound the expected regret of the exploit part, we will use our Full-Information algorithm "box":

- Input:  $(c_{i,a_1}, \dots, c_{i,a_k})$ .
- Performance  $\forall a \in [K]$ :

$$\sum_{i=1}^{\frac{T}{B}} \sum_{a'} P_i^{FI}(a') c_i(a') \leq \sum_{i=1}^{\frac{T}{B}} c_i(a) + \text{Regret}\left(\frac{T}{B}, K\right) \quad (4)$$

Important:  $c_i(a')$  and  $c_i(a)$  are random variables.

Now let us calculate the expected performance  $\forall a \in [K]$ :

$$\mathbb{E}\left[\sum_{i=1}^{\frac{T}{B}} \sum_{a'} P_i^{FI}(a') c_i(a')\right] \leq \mathbb{E}\left[\sum_{i=1}^{\frac{T}{B}} c_i(a)\right] + \text{Regret}\left(\frac{T}{B}, K\right) \stackrel{(2)}{\implies} \quad (5)$$

$$\frac{1}{B} \mathbb{E}\left[\sum_{i=1}^{\frac{T}{B}} \sum_{a'} P_i^{FI}(a') L_i(a')\right] \leq \frac{1}{B} \mathbb{E}\left[\sum_{i=1}^{\frac{T}{B}} L_i(a)\right] + \text{Regret}\left(\frac{T}{B}, K\right) \quad (6)$$

$$\mathbb{E}\left[\sum_{i=1}^{\frac{T}{B}} \sum_{a'} P_i^{FI}(a') L_i(a')\right] \leq \mathbb{E}\left[\sum_{i=1}^{\frac{T}{B}} L_i(a)\right] + B \cdot \text{Regret}\left(\frac{T}{B}, K\right) \quad (7)$$

In lecture 5 we saw RWM algorithm - a Full-Information algorithm with

$$\text{regret}\left(\frac{T}{B}, K\right) = \sqrt{\frac{T}{B} \log K}.$$

Let us use this algorithm with:  $B = T^{\frac{1}{3}} K^{\frac{2}{3}} \log^{-\frac{1}{3}} K$  and get:

$$\mathbb{E}[\text{Regret}_{\text{OnlineMAB}}] \leq 2T^{\frac{2}{3}} K^{\frac{1}{3}} \log^{\frac{1}{3}} K, \quad (8)$$

as claimed.  $\square$

### 3.4 Many Experts and Few Actions

Assume that we have many experts - marked by  $N$ , and few actions - marked by  $K$ , i.e.,  $K \ll N$ . For example a binary classification problem has  $K = 2$ , and can have many "experts" (=hypotheses).

Let us analyze the above algorithm for this scenario.

The Exploration cost is  $K \frac{T}{B}$ , therefore:

$$\mathbb{E}[\text{Regret}] \leq B \cdot R\left(\frac{T}{B}, N\right) + K \frac{T}{B} = 2\sqrt{TB \log N} + K \frac{T}{B} \quad (9)$$

Choose  $B = T^{\frac{1}{3}} K^{\frac{1}{2}} \log^{-\frac{1}{3}} N$  and get:

$$\mathbb{E}[\text{Regret}] \leq 2T^{\frac{2}{3}} K^{\frac{1}{2}} \log^{\frac{1}{3}} N, \quad (10)$$

Which has only a logarithmic dependence on  $N$ .

## 4 Conditional Expectation and Importance Sampling

### 4.1 Conditional Expectation

Let  $X, Y$  be real random variables, then:

- $\mathbb{E}[Y]$  is scalar.
- $\mathbb{E}[Y | X = 3]$  is scalar.
- $\mathbb{E}[Y | X = x]$  is a function of  $x$ , i.e.  $f(x)$ .

Assume  $X$  and  $Y$  are independent and let  $Z = X + Y$ , then:

- $\mathbb{E}[Z] = \mathbb{E}[X] + \mathbb{E}[Y]$
- $\mathbb{E}[Z | X = x] = \mathbb{E}[X | X = x] + \mathbb{E}[Y | X = x] = x + \mathbb{E}[Y]$

### 4.2 Importance Sampling

We sample  $X$  from a distribution  $D$ , and want to 'transform' it to a new distribution  $Q$ , let  $Y \triangleq X \frac{Q(X)}{D(X)}$  and we get:

$$\mathbb{E}_{X \sim D}[Y] = \sum_x x \frac{Q(x)}{D(x)} D(x) = \sum_x x Q(x) = \mathbb{E}_{X \sim Q}[X]$$

## 5 EXP3: algorithm for minimizing the expected regret

### 5.1 EXP3: Algorithm

EXP3 is a similar algorithm to WRM, with the following differences:

- Let us go back to reward terms:  $g_t(a) \in [0, 1]$ , where  $a_t$  is the selected action.
- Instead of full information, we now observe  $g_t(a_t)$  only.
- We mix 'uniform' distribution to  $p_t$ .
- We construct  $\hat{g}_t(a) = \frac{g_t(a_t)}{p_t(a_t)}$ , if  $a = a_t$ , and 0 otherwise.

```

EXP3( $\eta$ ),  $\eta \in [0, 1]$ 

initialize  $\omega_1(a) = 1, p_1(a) = \frac{1}{K} \quad \forall a \in [K]$ 
for  $t = 1, \dots, T$ :
  Select action  $a_t = a$  with probability  $p_t(a)$ 
   $W_t = \sum_a \omega_t(a)$ 
   $p_t(a) = (1 - \eta) \frac{\omega_t(a)}{W_t} + \eta \frac{1}{K} \quad \forall a \in [K]$ 
  Observe  $g_t(a_t)$ 
  Update  $\forall a \in [K]$  :
     $\hat{g}_t(a) = \begin{cases} \frac{g_t(a_t)}{p_t(a_t)}, & \text{if } a = a_t \\ 0, & \text{otherwise.} \end{cases}$ 
     $\omega_{t+1}(a) = \omega_t(a) e^{\frac{\eta \hat{g}_t(a)}{K}}$ 
end for

```

Figure 3: Code for EXP3 algorithm

### 5.2 EXP3: Regret Analysis

Definitions:

- $G(a) = \sum_{t=1}^T g_t(a)$
- $G^* = \max_a \{G(a)\}$
- $G(\text{EXP3}) = \sum_{t=1}^T g_t(a_t)$ , where  $g_t(a_t)$  depends on the the entire history.

Assumptions:

- $g_t(a) \in [0, 1]$
- $\exists \bar{G}$  such that  $\bar{G} \geq G^*$

**Lemma 3**  $G(\text{EXP3}) \geq (1 - \eta) \sum_{t=1}^T \hat{g}_t(a) - \frac{K \ln K}{\eta} - \frac{\eta}{K} \sum_{t=1}^T \sum_{a'} \hat{g}_t(a') \quad \forall a \in [K]$

**Proof:** Since  $\forall a \in [K]$ :

$$\hat{g}_t(a) \equiv \begin{cases} \frac{g_t(a_t)}{p_t(a_t)}, & \text{if } a = a_t \\ 0, & \text{otherwise} \end{cases} \leq \frac{g_t(a_t)}{p_t(a_t)} \leq \frac{1}{p_t(a_t)} \leq \frac{K}{\eta}. \quad (11)$$

Also:

$$\mathbb{E}[\hat{g}_t(a)] \equiv \sum_a p_t(a) \hat{g}_t(a) = p_t(a_t) \frac{g_t(a_t)}{p_t(a_t)} = g_t(a_t). \quad (12)$$

And:

$$\mathbb{E}[(\hat{g}_t(a))^2] \equiv \sum_a p_t(a) (\hat{g}_t(a))^2 = g_t(a_t) \hat{g}_t(a_t) \leq \hat{g}_t(a_t) = \sum_a \hat{g}_t(a) \quad (13)$$

Recall that  $\omega_{t+1}(a) = \omega_t(a) e^{\frac{\eta \hat{g}_t(a)}{K}}$ , Therefore:

$$\omega_{T+1}(a) = \omega_T(a) e^{\frac{\eta \hat{g}_T(a)}{K}} = \dots = \omega_1(a) e^{\frac{\eta}{K} \sum_{t=1}^T \hat{g}_t(a)} \quad (14)$$

Hence:

$$W_{T+1} \equiv \sum_a \omega_{T+1}(a) \geq \omega_{T+1}(a^*) \stackrel{(14)}{=} \omega_1(a^*) e^{\frac{\eta}{K} \sum_{t=1}^T \hat{g}_t(a^*)} \quad (15)$$

Also:

$$W_1 \equiv \sum_a \omega_1(a) = \sum_a 1 = K \quad (16)$$

Therefore:

$$\ln \frac{W_{T+1}}{W_1} \stackrel{(15),(16)}{\geq} \ln \frac{e^{\frac{\eta}{K} \sum_{t=1}^T \hat{g}_t(a^*)}}{K} = \frac{\eta}{K} \sum_{t=1}^T \hat{g}_t(a^*) - \ln K \quad (17)$$

Note that:

$$e^{\frac{\eta}{K} \hat{g}_t(a)} \stackrel{e^z \leq 1+z+z^2, \forall z \leq 1}{\leq} 1 + \frac{\eta}{K} \hat{g}_t(a) + \left(\frac{\eta}{K} \hat{g}_t(a)\right)^2 \quad (18)$$

Also, by definition of  $p_t(a)$ :  $\frac{\omega_t(a)}{W_t} = \frac{p_t(a) - \eta/K}{1-\eta}$ . Hence:

$$\begin{aligned} \frac{W_{t+1}}{W_t} &= \frac{\sum_a \omega_t(a) e^{\frac{\eta \hat{g}_t(a)}{K}}}{W_t} = \sum_a \frac{(p_t(a) - \eta/K) e^{\frac{\eta \hat{g}_t(a)}{K}}}{1-\eta} \\ &\stackrel{(18)}{\leq} \sum_a \frac{(p_t(a) - \eta/K) \left(1 + \frac{\eta}{K} \hat{g}_t(a) + \left(\frac{\eta}{K} \hat{g}_t(a)\right)^2\right)}{1-\eta} \\ &\leq \sum_a \frac{(p_t(a) - \eta/K)}{1-\eta} + \sum_a \frac{p_t(a) \frac{\eta}{K} \hat{g}_t(a)}{1-\eta} + \sum_a \frac{p_t(a) \left(\frac{\eta}{K} \hat{g}_t(a)\right)^2}{1-\eta} \\ &\stackrel{(12),(13)}{\leq} 1 + \frac{\eta}{K(1-\eta)} g_t(a_t) + \frac{\eta^2}{K^2(1-\eta)} \sum_a \hat{g}_t(a) \end{aligned} \quad (19)$$

Now, since  $\ln(x) \leq x - 1$ :

$$\ln \frac{W_{t+1}}{W_t} \stackrel{(19)}{\leq} \frac{\eta}{K(1-\eta)} g_t(a_t) + \frac{\eta^2}{K^2(1-\eta)} \sum_a \hat{g}_t(a) \quad (20)$$

Then:

$$\begin{aligned} \ln \frac{W_{T+1}}{W_1} &= \ln \prod_{t=1}^T \frac{W_{t+1}}{W_t} = \sum_{t=1}^T \ln \frac{W_{t+1}}{W_t} \\ &\stackrel{(20)}{\leq} \sum_{t=1}^T \left( \frac{\eta}{K(1-\eta)} g_t(a_t) + \frac{\eta^2}{K^2(1-\eta)} \sum_a \hat{g}_t(a) \right) \end{aligned} \quad (21)$$

Putting together (17) and (21) and get:

$$\frac{\eta}{K} \sum_{t=1}^T \hat{g}_t(a^*) - \ln K \leq \ln \frac{W_{T+1}}{W_1} \leq \sum_{t=1}^T \left( \frac{\eta}{K(1-\eta)} g_t(a_t) + \frac{\eta^2}{K^2(1-\eta)} \sum_a \hat{g}_t(a) \right) \quad (22)$$

Which equivalents to:

$$(1-\eta) \sum_{t=1}^T \hat{g}_t(a^*) - \frac{K \ln K}{\eta} - \frac{\eta}{K} \sum_a \sum_{t=1}^T \hat{g}_t(a) \leq \sum_{t=1}^T g_t(a_t) \equiv G(\text{EXP3}), \quad (23)$$

as claimed.  $\square$

We are now in a position to prove the main result:

**Theorem 4**  $\mathbb{E}[\text{Regret}] \equiv G^* - \mathbb{E}[G(\text{EXP3})] \leq 2\eta G^* + \frac{K \ln K}{\eta}$

and for  $\eta = \sqrt{\frac{K \ln K}{2G}}$  :  $\mathbb{E}[\text{Regret}] \leq 2\sqrt{2GK \ln K}$ .

**Proof:** Let us take an expectation of **Lemma 3** expression and get:

$$\begin{aligned} \mathbb{E}[G(\text{EXP3})] &\geq (1-\eta) \sum_{t=1}^T \mathbb{E}[\hat{g}_t(a)] - \frac{K \ln K}{\eta} - \frac{\eta}{K} \sum_{t=1}^T \sum_{a'} \mathbb{E}[\hat{g}_t(a')], \quad \forall a \in [K] \\ &\stackrel{(12)}{=} (1-\eta)G(a) - \frac{K \ln K}{\eta} - \frac{\eta}{K} \sum_{a'} G(a'), \quad \forall a \in [K] \end{aligned} \quad (24)$$

Therefore:

$$\mathbb{E}[G(\text{EXP3})] \geq (1-\eta)G^* - \frac{K \ln K}{\eta} - \eta G^*. \quad (25)$$

and:

$$G^* - \mathbb{E}[G(\text{EXP3})] \leq 2\eta G^* + \frac{K \ln K}{\eta} \quad (26)$$

By definition,  $\hat{G}^* \leq \bar{G}$ , hence, for  $\eta = \sqrt{\frac{K \ln K}{2G}}$  :

$$\mathbb{E}[\text{Regret}] \leq 2\sqrt{2GK \ln K}, \quad (27)$$

which concludes the proof.  $\square$

## 6 EXP-IX: algorithm for high probability regret

### 6.1 EXP-IX: Algorithm

EXP3 optimizes the expected regret, but it has a large variance. However, EXP-IX, insures minimum regret with high probability. Main idea - change the estimator:

$$\tilde{l}_t(a) \triangleq \frac{l_t(a)}{p_t(a) + \gamma} \mathbb{1}\{a = a_t\}, \quad (28)$$

where  $l_t(a) \in [0, 1]$  is the loss in time  $t$  of action  $a$  (returning to loss terms). Note that for  $\gamma > 0$  the estimator above is biased, in contrast to EXP3.

```

EXP-IX( $\eta, \gamma$ ),    $\eta \in [0, 1], \quad \gamma \in [0, 1],$ 

  initialize  $\omega_1(a) = 1, p_1(a) \frac{1}{K} \quad \forall a \in [K]$ 
  for  $t = 1, \dots, T$ :
    Select action  $a_t = a$  with probability  $p_t(a)$ 
     $W_t = \sum_a \omega_t(a)$ 
     $p_t(a) = (1 - \eta) \frac{\omega_t(a)}{W_t} + \eta \frac{1}{K} \quad \forall a \in [K]$ 
    Observe  $l_t(a_t)$ 
    Update  $\forall a \in [K]$  :
       $\tilde{l}_t(a) \triangleq \frac{l_t(a)}{p_t(a) + \gamma} \mathbb{1}\{a = a_t\}$ 
       $\omega_{t+1}(a) = \omega_t(a) e^{-\frac{\eta \tilde{l}_t(a)}{K}}$ 
  end for

```

Figure 4: Code for EXP-IX algorithm

### 6.2 EXP-IX: Regret Analysis

**Lemma 5**  $\mathbb{P}\left(\sum_{t=1}^T \left(\tilde{l}_t(a) - l_t(a)\right) \leq \frac{\ln(K/\delta)}{2\gamma}\right) > 1 - \delta, \quad \forall a \in [K].$

**Proof:** Note that:  $\mathbb{E}\left[\tilde{l}_t(a)\right] < l_t(a), \quad \forall \gamma \in [0, 1]$  since:

$$\mathbb{E}\left[\tilde{l}_t(a)\right] \equiv \mathbb{E}\left[\frac{l_t(a)}{p_t(a) + \gamma} \mathbb{1}\{a = a_t\}\right] = p_t(a) \frac{l_t(a)}{p_t(a) + \gamma} < l_t(a), \quad \forall \gamma > 0. \quad (29)$$

Now let  $\beta \triangleq 2\gamma$  and note that the following holds:

$$\begin{aligned} \tilde{l}_t(a) &\equiv \frac{l_t(a)}{p_t(a) + \gamma} \mathbb{1}\{a = a_t\} \leq \frac{l_t(a)}{p_t(a) + \gamma} \\ &= \frac{1}{2\gamma} \frac{2\gamma l_t(a)/p_t(a)}{1 + \gamma l_t(a)/p_t(a)} \stackrel{\frac{z}{1+z/2} \leq \ln(1+z), \forall z \geq 0}{\leq} \frac{1}{\beta} \ln\left(1 + \beta \hat{l}_t(a)\right) \end{aligned} \quad (30)$$

Then:

$$\mathbb{E}\left[e^{\beta \tilde{l}_t(a)} | \mathcal{H}_{t-1}\right] \stackrel{(30)}{\leq} \mathbb{E}\left[1 + \beta \hat{l}_t(a) | \mathcal{H}_{t-1}\right] \stackrel{(29)}{\leq} 1 + \beta l_t(a) \stackrel{1+z \leq e^z}{\leq} e^{\beta l_t(a)} \quad (31)$$

Let  $Z_t \triangleq \exp\left(\beta \sum_{t=1}^T (\tilde{l}_t(a) - l_t(a))\right)$ , we proved that  $\mathbb{E}[Z_t | \mathcal{H}_{t-1}] \leq 1$ .

Therefore from Markov inequality:

$$\mathbb{P}\left(\sum_{t=1}^T (\tilde{l}_t(a) - l_t(a)) > \lambda\right) \leq \mathbb{E}\left[\exp\left(\beta \sum_{t=1}^T (\tilde{l}_t(a) - l_t(a))\right)\right] e^{-\lambda\beta} \leq e^{-\lambda\beta}. \quad (32)$$

Choose  $\lambda = \frac{1}{\beta} \ln \frac{K}{\delta} \Leftrightarrow e^{-\lambda\beta} = \frac{\delta}{K}$  and get:

$$\mathbb{P}\left(\sum_{t=1}^T (\tilde{l}_t(a) - l_t(a)) > \frac{\ln(K/\delta)}{2\gamma}\right) \leq \delta, \quad (33)$$

as claimed.  $\square$

We are now in a position to prove the main result:

**Theorem 6**  $\mathbb{P}\left(\text{Regret} > O\left(\sqrt{TK \ln K} + \sqrt{TK} \frac{\ln \frac{K}{\delta}}{\sqrt{\ln K}}\right)\right) \leq \delta$ .

**Proof:** We can use EXP3 Lemma:

$$\sum_{t=1}^T \left(\sum_{a'} p_t(a') \tilde{l}_t(a') - \tilde{l}_t(a)\right) \leq \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{t=1}^T \sum_{a'} p_t(a') (\tilde{l}_t(a'))^2 \quad (34)$$

Where:

$$\begin{aligned} \sum_a p_t(a) \tilde{l}_t(a) &= \sum_a \frac{l_t(a) (p_t(a) + \gamma)}{p_t(a) + \gamma} \mathbb{1}\{a = a_t\} - \gamma \sum_a \frac{l_t(a)}{p_t(a) + \gamma} \mathbb{1}\{a = a_t\} \\ &= l_t(a_t) - \gamma \sum_a \tilde{l}_t(a) \end{aligned} \quad (35)$$

And:

$$\sum_a p_t(a) (\tilde{l}_t(a))^2 \leq \sum_a \tilde{l}_t(a) \quad (36)$$

Putting together and get  $\forall a \in [K]$ :

$$\sum_{t=1}^T l_t(a_t) - \gamma \sum_{t=1}^T \sum_{a'} \tilde{l}_t(a') - \sum_{t=1}^T \tilde{l}_t(a) \leq \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{t=1}^T \sum_{a'} \tilde{l}_t(a') \quad (37)$$

Which equivalents to:

$$\sum_{t=1}^T (l_t(a_t) - l_t(a)) \leq \sum_{t=1}^T (\tilde{l}_t(a) - l_t(a)) + \frac{\ln K}{\eta} + \left(\frac{\eta}{2} + \gamma\right) \sum_{t=1}^T \sum_{a'} \tilde{l}_t(a') \quad (38)$$

Then, from **Lemma 5** we get:

$$\mathbb{P}\left(\text{Regret} \leq \frac{\ln \frac{2K}{\delta}}{2\gamma} + \frac{\ln K}{\eta} + \left(\frac{\eta}{2} + \gamma\right) \underbrace{\sum_{t=1}^T \sum_{a'} l_t(a')}_{\text{TK}} + K \frac{\ln \frac{2K}{\delta}}{2\gamma} \left(\frac{\eta}{2} + \gamma\right)\right) > 1 - \delta \quad (39)$$

Choose  $\gamma = \frac{\eta}{2}$ ,  $\eta = \sqrt{\frac{\ln K}{TK}}$ , and get:

$$\mathbb{P}\left(\text{Regret} \leq O\left(\sqrt{TK \ln K} + \sqrt{TK} \frac{\ln \frac{K}{\delta}}{\sqrt{\ln K}}\right)\right) > 1 - \delta, \quad (40)$$

which concludes the proof.  $\square$

## 7 References

1. RWM algorithm - Lecture 5 notes.
2. EXP3 algorithm - P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32:48–77, 2002.
3. EXP-IX algorithm - G. Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *Advances in Neural Information Processing Systems*, pp. 3168–3176, 2015.