

# Machine Learning Theory 2022

## Lecture 9

**Wouter M. Koolen**

Download these slides now from [elo.mastermath.nl](http://elo.mastermath.nl)!

Bandit problems

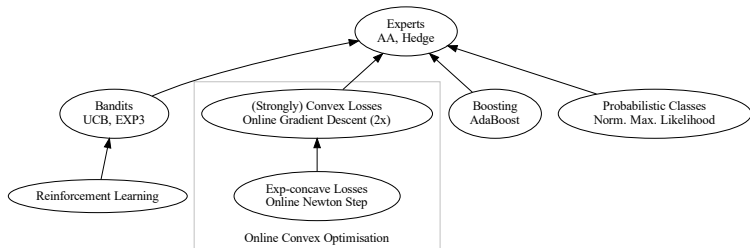
- ▶ Adversarial bandit Setting
- ▶ EXP3 Algorithm and analysis
- ▶ Stochastic bandit Setting
- ▶ UCB Algorithm and analysis



homework roulette  
in the break

## Recap and Bandit Setting

# Overview of Second Half of Course



Material: course notes and selection of sources on MLT website.

# Recap: Setting

## Protocol (Dot Loss Game)

- ▶ For  $t = 1, 2, \dots$ 
  - ▶ Learner chooses a distribution  $\mathbf{w}_t \in \Delta_K$  on  $K$  “experts”.
  - ▶ Adversary reveals loss vector  $\ell_t \in [0, 1]^K$ .
  - ▶ Learner’s loss is the **dot loss**  $\mathbf{w}_t^\top \ell_t = \sum_{k=1}^K w_t^k \ell_t^k$

## Objective

*Regret* after  $T$  rounds:

$$R_T = \underbrace{\sum_{t=1}^T \mathbf{w}_t^\top \ell_t}_{\text{Learner's dot loss}} - \underbrace{\min_k \sum_{t=1}^T \ell_t^k}_{\text{loss of best expert}}$$

A **good learner** has **small regret**, i.e. it approximately behaves as if it knows the *best expert*.

## Recap: Method and Result

### Definition (Hedge Algorithm)

The *Hedge algorithm* with learning rate  $\eta$  plays weights in round  $t$ :

$$w_t^k = \frac{e^{-\eta \sum_{s=1}^{t-1} \ell_s^k}}{\sum_{j=1}^K e^{-\eta \sum_{s=1}^{t-1} \ell_s^j}}. \quad (\text{Hedge})$$

or, equivalently,  $w_1^k = \frac{1}{K}$  and

$$w_{t+1}^k = \frac{w_t^k e^{-\eta \ell_t^k}}{\sum_{j=1}^K w_t^j e^{-\eta \ell_t^j}} \quad (\text{Hedge, incremental})$$

### Theorem (Hedge Regret Bound)

The regret of Hedge with learning rate  $\eta = \sqrt{\frac{8 \ln K}{T}}$  is at most

$$R_T \leq \sqrt{T/2 \ln K}.$$

# “Bandits”: a (minor?) change of feedback

Learner picks  $w_t \in \Delta_K$   
Environment determines loss  $\ell_t \in [0, 1]^K$   
Learner sees **full**  $\ell_t$



Learner picks action  $I_t \in [K]$  (possibly at random)  
Environment determines loss  $\ell_t \in [0, 1]^K$   
Learner sees **only**  $\ell_t^{I_t}$

**Radical upgrade:** Learner actively controls **which data** are collected.

## Applications

- ▶ Clinical trials (round=patient, action=treatment)
- ▶ Advertising (round=visitor, action=serving specific ad)
- ▶ Radio channel selection (wifi)
- ▶ ...

# Main Questions

- ▶ How difficult is it to learn from **partial observations**?
- ▶ How should learning algorithms be (re)designed?
  - ▶ Obtaining information requires executing sub-optimal actions
  - ▶ Exploration/Exploitation trade-off
- ▶ What is the effect of the environment model?
  - ▶ Adversarial
  - ▶ Stochastic

Different techniques, different complexity (regret rate)

# Two Brilliant Ideas



- ▶ Importance Weighted Loss Estimates
- ▶ Optimism in Face of Uncertainty



# Adversarial Bandits

# Main Questions

How difficult is it to learn from **partial observations**?

# The setup

## Protocol ( $K$ -armed adversarial bandit)

- ▶ Adversary hides  $\ell_t^k \in [0, 1]$  for all  $t \leq T, k \leq K$ .
- ▶ For  $t = 1, 2, \dots, T$ 
  - ▶ Learner picks arm  $I_t$  (typically by sampling  $I_t \sim \mathbf{w}_t$ )
  - ▶ Learner observes and incurs loss  $\ell_t^{I_t}$

## Objective:

Expected regret w.r.t. best arm after  $T$  rounds:

$$\mathbb{E}[R_T] = \mathbb{E}_{I_1 \dots I_T} \left[ \sum_{t=1}^T \ell_t^{I_t} \right] - \min_k \sum_{t=1}^T \ell_t^k$$

# Outline of this part

We will prove the following result:

## Theorem (Main Adversarial Bandit Result)

*There is an algorithm with regret*

$$\mathbb{E}[R_T] \leq C \cdot \sqrt{TK \ln K}$$

Ingredients:

- ▶ Importance weighted estimates
- ▶ Reduction to AA
- ▶ Tweaks to AA analysis

# Importance Weighted Loss Estimates

- ▶ Opponent fixed  $\ell_t$ .
- ▶ We draw  $I_t \sim w_t$ .
- ▶ We see  $\ell_t^{I_t}$ .

We only see **one entry** of  $\ell_t$ . Can we still **estimate** the full  $\ell_t$ ?

## Definition (Loss Estimate)

The *importance weighted loss estimate* is  $\hat{\ell}_t$  with entries  $\hat{\ell}_t^k := \frac{\ell_t^{I_t}}{w_t^{I_t}} \mathbf{1}_{I_t=k}$ .

## Example

Say  $K = 4$ ,  $w_t = (0.1, 0.2, 0.3, 0.4)$ ,  $\ell_t = (0.6, 0.7, 0.8, 0.9)$  and sampling from  $w_t$  gives  $I_t = 3$ . Then we see  $\ell_t^{I_t} = \ell_t^3 = 0.8$  and form the estimate

$$\hat{\ell}_t = \begin{pmatrix} 0 \\ 0 \\ 0.8/0.3 = 2.66\dots \\ 0 \end{pmatrix}.$$

# Importance Weighted Loss Estimates

## Definition (Loss Estimate)

We pick  $\hat{\ell}_t$  with  $\hat{\ell}_t^k = \frac{\ell_t^{l_t}}{w_t^{l_t}} \mathbf{1}_{l_t=k}$ .

## Lemma (Unbiased Estimator)

$$\mathbb{E}_{l_t \sim w_t}[\hat{\ell}_t] = \ell_t.$$

### Proof.

For each  $k$

$$\mathbb{E}_{l_t \sim w_t}[\hat{\ell}_t^k] = \sum_{l_t=1}^K w_t^{l_t} \frac{\ell_t^{l_t}}{w_t^{l_t}} \mathbf{1}_{l_t=k} = \sum_{l_t=1}^K \ell_t^{l_t} \mathbf{1}_{l_t=k} = \ell_t^k$$



## Corollary

$$\mathbb{E}_{l_t \sim w_t}[w_t^\top \hat{\ell}_t] = w_t^\top \ell_t = \mathbb{E}_{l_t \sim w_t}[\ell_t^{l_t}].$$

## EXP3: AA+scaling+estimation

Slogan: EXP3 is AA applied to  $\eta$ -scaled importance weighted losses.

### Definition (EXP3 Algorithm)

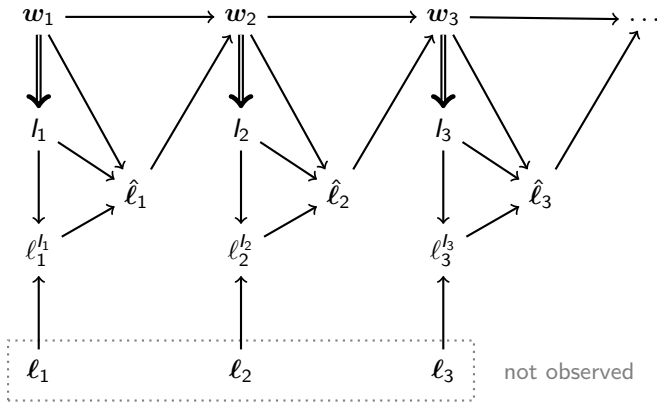
The *EXP3 algorithm* with *learning rate*  $\eta > 0$  plays weights in round  $t$ :

$$w_t^k = \frac{e^{-\eta \sum_{s=1}^{t-1} \hat{\ell}_s^k}}{\sum_{j=1}^K e^{-\eta \sum_{s=1}^{t-1} \hat{\ell}_s^j}}. \quad (\text{EXP3})$$

or, equivalently,  $w_1^k = \frac{1}{K}$  and

$$w_{t+1}^k = \frac{w_t^k e^{-\eta \hat{\ell}_t^k}}{\sum_{j=1}^K w_t^j e^{-\eta \hat{\ell}_t^j}} \quad (\text{EXP3, incremental})$$

# Dependency structure





## EXP3 Analysis: dot/mix loss

We would like to use the dot/mix loss relationship. For Hedge the losses  $\ell_t$  are **bounded**, and we could use Hoeffding's Inequality. But for EXP3 the importance weighted loss estimates  $\hat{\ell}_t$  are **not bounded** above. We need another relation.

### Lemma

For losses  $\hat{\ell}_t^k \geq 0$  and learning rate  $\eta > 0$ ,

$$\underbrace{\sum_{k=1}^K w_t^k \hat{\ell}_t^k}_{\text{dot loss on } \hat{\ell}_t} \leq \underbrace{-\frac{1}{\eta} \ln \left( \sum_{k=1}^K w_t^k e^{-\eta \hat{\ell}_t^k} \right)}_{\text{(scaled) mix loss on } \eta \hat{\ell}_t} + \underbrace{\frac{\eta}{2} \sum_{k=1}^K w_t^k (\hat{\ell}_t^k)^2}_{\text{overhead}} \quad (1)$$

## EXP3 Analysis: dot/mix loss

Proof.

For  $x \geq 0$ , we have  $e^{-x} \leq 1 - x + x^2/2$ . Hence

$$\begin{aligned} -\ln \left( \sum_{k=1}^K w_t^k e^{-\eta \hat{\ell}_t^k} \right) &\geq -\ln \left( \sum_{k=1}^K w_t^k (1 - \eta \hat{\ell}_t^k + (\eta \hat{\ell}_t^k)^2/2) \right) \\ &= -\ln \left( 1 - \eta \sum_{k=1}^K w_t^k \hat{\ell}_t^k + \frac{\eta^2}{2} \sum_{k=1}^K w_t^k (\hat{\ell}_t^k)^2 \right) \\ &\geq \eta \sum_{k=1}^K w_t^k \hat{\ell}_t^k - \frac{\eta^2}{2} \sum_{k=1}^K w_t^k (\hat{\ell}_t^k)^2 \end{aligned}$$

Dividing by  $\eta > 0$  and moving the rightmost term over gives the lemma. □

## EXP3 Analysis: overhead term

Let's study that overhead term in expectation

### Lemma

*In round  $t$ , for the importance weighted loss estimator  $\hat{\ell}_t$*

$$\mathbb{E}_{I_t \sim \mathbf{w}_t} \left[ \sum_{k=1}^K w_t^k (\hat{\ell}_t^k)^2 \right] \leq K \quad (2)$$

### Proof.

By definition of the importance weighted loss estimator

$$\mathbb{E}_{I_t \sim \mathbf{w}_t} \left[ \sum_{k=1}^K w_t^k (\hat{\ell}_t^k)^2 \right] = \sum_{I_t=1}^K w_t^{I_t} \sum_{k=1}^K w_t^k \left( \frac{\ell_t^{I_t}}{w_t^{I_t}} \mathbf{1}_{I_t=k} \right)^2$$

Only the diagonal  $I_t = k$  contributes, and the loss is bounded  $\ell_t^{I_t} \in [0, 1]$ ,  
so

$$= \sum_{I_t=1}^K w_t^{I_t} w_t^{I_t} \frac{(\ell_t^{I_t})^2}{(w_t^{I_t})^2} = \sum_{I_t=1}^K (\ell_t^{I_t})^2 \leq K.$$

# EXP3 Regret Bound

## Theorem

*The expected regret of EXP3 with learning rate  $\eta > 0$  is*

$$\mathbb{E}_{I_1 \dots I_T} \left[ \sum_{t=1}^T \ell_t^{I_t} \right] - \min_k \sum_{t=1}^T \ell_t^k \leq \frac{\ln K}{\eta} + \frac{TK\eta}{2}$$

## Corollary

*The expected regret of EXP3 with learning rate  $\eta = \sqrt{\frac{2 \ln K}{TK}}$  is*

$$\mathbb{E}_{I_1 \dots I_T} \left[ \sum_{t=1}^T \ell_t^{I_t} \right] - \min_k \sum_{t=1}^T \ell_t^k \leq \sqrt{2TK \ln K}$$

## EXP3 Analysis

### Proof.

Sum the mix/dot loss inequality (1) over rounds

$$\sum_{t=1}^T \sum_{k=1}^K w_t^k \hat{\ell}_t^k \leq \sum_{t=1}^T -\frac{1}{\eta} \ln \left( \sum_{k=1}^K w_t^k e^{-\eta \hat{\ell}_t^k} \right) + \sum_{t=1}^T \frac{\eta}{2} \sum_{k=1}^K w_t^k (\hat{\ell}_t^k)^2$$

Apply the AA regret bound (Lecture 8) to the middle sum

$$\sum_{t=1}^T \sum_{k=1}^K w_t^k \hat{\ell}_t^k \leq \min_k \sum_{t=1}^T \hat{\ell}_t^k + \frac{\ln K}{\eta} + \sum_{t=1}^T \frac{\eta}{2} \sum_{k=1}^K w_t^k (\hat{\ell}_t^k)^2$$

Take expectations, and pull the min out using Jensen's Inequality

$$\mathbb{E}_{I_1 \dots I_T} \left[ \sum_{t=1}^T \sum_{k=1}^K w_t^k \hat{\ell}_t^k \right] \leq \min_k \mathbb{E}_{I_1 \dots I_T} \left[ \sum_{t=1}^T \hat{\ell}_t^k \right] + \frac{\ln K}{\eta} + \mathbb{E}_{I_1 \dots I_T} \left[ \sum_{t=1}^T \frac{\eta}{2} \sum_{k=1}^K w_t^k (\hat{\ell}_t^k)^2 \right]$$

Use unbiasedness and our expected overhead bound (2) to conclude

$$\mathbb{E}_{I_1 \dots I_T} \left[ \sum_{t=1}^T \ell_t^k \right] \leq \min_k \sum_{t=1}^T \ell_t^k + \frac{\ln K}{\eta} + T \frac{\eta}{2} K.$$

# Conclusion of Adversarial Bandits part

An algorithm that can learn from **partial information** even with **adversarially determined losses**.

Observations:

- ▶ Efficient: run time is order  $K$  per round.
- ▶ Regret of EXP3 is  $\sqrt{TK \ln K}$  compared to  $\sqrt{T \ln K}$  for Hedge in full information setting.
- ▶ For  $\sqrt{KT}$  lower bound see reading material.
- ▶ Exploration/exploitation. Unsamped arms get 0 estimated loss. So eventually they will get sampled.

# Stochastic Bandits

# Main Questions

How difficult is it to learn from **partial observations** if we assume the **environment is stochastic**?

Are we back in statistical learning?

Not quite:

- ▶ Yes: statistical model for environment
- ▶ No (Major): Learner actively controls **which data** are collected
- ▶ No (Minor): **sequential** evaluation

A completely different style of algorithm.



# Setting

## Protocol ( $K$ -armed stochastic bandit)

- ▶ Environment: distributions  $(\nu_1, \dots, \nu_K)$  of arm rewards
- ▶ For  $t = 1, 2, \dots, T$ 
  - ▶ Learner picks arm  $I_t$
  - ▶ Learner observes and receives reward  $X_t \sim \nu_{I_t}$

## Definition (Stochastic Bandit Notation)

The *mean reward* of arm  $k$  is  $\mu^k = \mathbb{E}_{X \sim \nu_k}[X]$ . The *best arm* is  $i^* = \arg \max_k \mu^k$ . The *sub-optimality gap* of arm  $i$  is  $\Delta_i = \mu^{i^*} - \mu^i$ .

## Objective

Pseudo-regret after  $T$  rounds:

$$\bar{R}_T = T\mu^{i^*} - \mathbb{E}_{I_1 \dots I_T} \left\{ \sum_{t=1}^T \mu^{I_t} \right\}$$

# Outline of this part

We will prove the following result:

## Theorem (Main Stochastic Bandit Result)

*There is an algorithm with pseudo-regret*

$$\bar{R}_T \leq C \left( \sum_{k \neq i^*}^K \frac{1}{\Delta_k} \right) \ln T + C'$$

# Idea

- ▶ For each arm, **estimate** its mean.
- ▶ **Empirical estimate** of mean of arm  $k$  after  $t$  rounds:

$$\hat{\mu}_t^k = \frac{\sum_{s=1}^t X_s 1_{I_s=k}}{N_t^k} \quad \text{where} \quad N_t^k = \sum_{s=1}^t 1_{I_s=k}$$

- ▶ Uncertainty quantification by means of a confidence interval

$$\text{LCB}_t^k := \hat{\mu}_t^k - \sqrt{\frac{2\alpha \ln(t+1)}{N_t^k}}$$
$$\text{UCB}_t^k := \hat{\mu}_t^k + \sqrt{\frac{2\alpha \ln(t+1)}{N_t^k}}$$

Claim: True mean  $\mu^k \in [\text{LCB}_t^k, \text{UCB}_t^k]$  with near-certainty (probability  $\approx 1$ ).

- ▶ Sample the arm of highest  $\text{UCB}_t$

# UCB Algorithm

## Definition (UCB Algorithm)

In round  $t$ , the UCB algorithm with parameter  $\alpha > 2$  samples arm

$$I_t := \arg \max_k \text{UCB}_{t-1}^k = \arg \max_k \hat{\mu}_{t-1}^k + \sqrt{\frac{2\alpha \ln(t)}{N_{t-1}^k}}$$

Optimism in face of uncertainty

Take the action of highest reward among any plausible bandit models.

# Where we are heading

We will show

## Theorem (UCB Regret Bound)

*UCB with  $\alpha > 2$  satisfies*

$$\bar{R}_T = \sum_{k=1}^K \mathbb{E}[N_T^k] \Delta_k \leq \left( \sum_{k \neq i^*}^K \frac{1}{\Delta_i} \right) 8\alpha \ln T + \frac{\alpha}{\alpha - 2} \sum_{k=1}^K \Delta_k$$

# UCB Analysis

Let  $i^* = \arg \max_k \mu^k$  be the index of the arm of highest mean.

If in some round  $t$  the algorithm samples a suboptimal arm,  $I_t = i \neq i^*$ , one of three things must be the case

- ▶ We have not sampled arm  $i$  often; its confidence width is still large
- ▶ Arm  $i$  is overestimated (its  $\text{LCB}_{t-1}^i$  is too high).
- ▶ Arm  $i^*$  is underestimated (its  $\text{UCB}_{t-1}^{i^*}$  is too low).

# UCB Analysis

## Lemma

If UCB samples  $I_t = i$  then either

- (a)  $\text{UCB}_{t-1}^{i*} \leq \mu^{i*}$
- (b)  $\text{LCB}_{t-1}^i > \mu_i$
- (c)  $N_{t-1}^i < \frac{8\alpha \ln t}{\Delta_i^2}$  where  $\Delta_i = \mu^{i*} - \mu^i$ .

## Proof.

Suppose not. Then

$$\begin{aligned} \text{UCB}_{t-1}^{i*} &\stackrel{(a)}{>} \mu^{i*} = \mu^i + \Delta_i \stackrel{(c)}{\geq} \mu^i + 2\sqrt{\frac{2\alpha \ln t}{N_{t-1}^i}} \\ &\stackrel{(b)}{\geq} \text{LCB}_{t-1}^i + 2\sqrt{\frac{2\alpha \ln t}{N_{t-1}^i}} \stackrel{\text{def.}}{=} \text{UCB}_{t-1}^i \end{aligned}$$

and this contradicts that UCB samples  $I_t = i$ .



# UCB Analysis

The pseudo-regret can be rewritten as

$$\bar{R}_T = T\mu^{i^*} - \mathbb{E} \left[ \sum_{t=1}^T \mu^{I_t} \right] = \sum_{k=1}^K \mathbb{E}[N_T^k] \Delta_k$$

It hence suffices to bound  $\mathbb{E}[N_T^i]$  for suboptimal  $i \neq i^*$ .

We will show for each  $k \neq i^*$

$$\mathbb{E}[N_T^k] \leq C \frac{8\alpha \ln T}{\Delta_k^2} + C'$$

Slogan: Sub-optimal arms are sampled logarithmically often



# UCB Analysis

$$\begin{aligned}\mathbb{E}[N_T^i] &= \mathbb{E}\left[\sum_{t=1}^T 1_{I_t=i}\right] = \sum_{t=1}^T \mathbb{P}\{I_t = i\} \\&= \sum_{t=1}^T \mathbb{P}\{I_t = i \text{ and } N_{t-1}^i < u\} + \sum_{t=1}^T \mathbb{P}\{I_t = i \text{ and } N_{t-1}^i \geq u\} \\&\leq u + \sum_{t=u+1}^T \mathbb{P}\{I_t = i \text{ and } N_{t-1}^i \geq u\} \\&\stackrel{\text{Lemma}}{\leq} u + \sum_{t=u+1}^T \mathbb{P}\left\{\text{UCB}_{t-1}^{i*} \leq \mu^{i*} \text{ or } \text{LCB}_{t-1}^i > \mu_i\right\} \\&\stackrel{\text{Union bd.}}{\leq} u + \sum_{t=u+1}^T \mathbb{P}\left\{\text{UCB}_{t-1}^{i*} \leq \mu^{i*}\right\} + \sum_{t=u+1}^T \mathbb{P}\left\{\text{LCB}_{t-1}^i > \mu_i\right\}\end{aligned}$$

where  $u = \lceil \frac{8\alpha \ln T}{\Delta_i^2} \rceil \leq \frac{8\alpha \ln T}{\Delta_i^2} + 1$ .

## Confidence bounds are valid

We need to control two similar deviation events. For the first we will show

$$\sum_{t=u+1}^T \mathbb{P} \left\{ \text{UCB}_{t-1}^{i*} \leq \mu^{i*} \right\} \leq \sum_{t=u+1}^T \frac{1}{t^{\alpha-1}} \leq \frac{1}{\alpha-2}.$$

Let  $\tilde{\mu}_n^i$  be the average of the first  $n$  samples from arm  $i$ , so that  $\hat{\mu}_{t-1}^i = \tilde{\mu}_{N_{t-1}^i}^i$ . Then

$$\begin{aligned} \mathbb{P} \left\{ \text{UCB}_{t-1}^{i*} \leq \mu^{i*} \right\} &\stackrel{\text{def}}{=} \mathbb{P} \left\{ \hat{\mu}_{t-1}^{i*} + \sqrt{\frac{2\alpha \ln(t)}{N_{t-1}^{i*}}} \leq \mu^{i*} \right\} \\ &= \mathbb{P} \left\{ \tilde{\mu}_{N_{t-1}^{i*}}^{i*} + \sqrt{\frac{2\alpha \ln(t)}{N_{t-1}^{i*}}} \leq \mu^{i*} \right\} \\ &\leq \mathbb{P} \left\{ \exists s \in \{1, \dots, t\} : \tilde{\mu}_s^{i*} + \sqrt{\frac{2\alpha \ln(t)}{s}} \leq \mu^{i*} \right\} \\ &\stackrel{\text{union bd}}{\leq} \sum_{s=1}^t \mathbb{P} \left\{ \tilde{\mu}_s^{i*} + \sqrt{\frac{2\alpha \ln(t)}{s}} \leq \mu^{i*} \right\} \\ &= \sum_{s=1}^t \mathbb{P} \left\{ \mu^{i*} - \tilde{\mu}_s^{i*} \geq \sqrt{\frac{2\alpha \ln(t)}{s}} \right\} \stackrel{\text{Chernoff}}{\leq} \sum_{s=1}^t e^{-s(\sqrt{\frac{2\alpha \ln(t)}{s}})^2/2} = \sum_{s=1}^t \frac{1}{t^\alpha} \end{aligned}$$

# Overall result

We proved

## Theorem (UCB Regret Bound)

*UCB with  $\alpha > 2$  satisfies*

$$\bar{R}_T = \sum_{k=1}^K \mathbb{E}[N_T^k] \Delta_k \leq \left( \sum_{k \neq i^*}^K \frac{1}{\Delta_i} \right) 8\alpha \ln T + \frac{\alpha}{\alpha - 2} \sum_{k=1}^K \Delta_k$$

# Conclusion of stochastic bandit part

An algorithm that can learn from **partial information** with **stochastic losses**.

Observations:

- ▶ Regret of UCB is  $\ln T$  whereas EXP3 is  $\sqrt{T}$ .
- ▶ Regret of UCB is **instance dependent** (through gaps  $\Delta_i$ )
- ▶ Exploration/exploitation mechanism: confidence intervals + optimism
- ▶ Matching lower bounds exist.