Recap

Chap 5

full feedback

0 **Deterministic oblivious adversary.** W.l.o.g., the entire "cost table" $(c_t(a): a \in [K], t \in [T])$ is chosen before round 1. The best arm is naturally defined as $\text{argmin}_{a\in[K]} \texttt{cost}(a)$, and regret is defined as

$$R(T) = \texttt{cost}(\texttt{ALG}) - \min_{a\in[K]} \texttt{cost}(a), \tag{5.1}$$

1 Random.

2 adaptive.

Algorithm 5.2: `Hedge` algorithm for online learning with experts

---

parameter : $\epsilon \in (0, \frac{1}{2})$

Initialize the weights as $w_1(a) = 1$ for each arm a.
For each round t:
  Let $p_t(a) = \frac{w_t(a)}{\sum_{a'=1}^{K} w_t(a')}$. expert ℓ
  Sample an arm a_t from distribution $p_t(\cdot)$.
  Observe cost $c_t(a)$ for each arm a.
  For each arm a, update its weight
  $w_{t+1}(a) = w_t(a) \cdot (1-\epsilon)^{c_t(a)}$. $\hat{G}(\ell)$.

Theorem 5.9. Consider online learning with experts, with a randomized-oblivious adversary. Assume the costs are independent across rounds. Assume upper bound (5.11) for some μ and σ known to the algorithm. Then `Hedge` with parameter $\epsilon = \sqrt{\ln K/(3T(\mu^2+\sigma^2))}$ has regret

$$\mathbb{E}[\texttt{cost}(\texttt{ALG}) - \texttt{cost}^*] < 2\sqrt{3} \cdot \sqrt{T(\mu^2+\sigma^2)\ln K}.$$

Thm 5.7

Theorem 6.1. Consider online learning with N experts. Consider adaptive adversary and regret $R(T)$ relative to the best-observed expert. Algorithm `Hedge` with parameter $\epsilon = \epsilon_U := \sqrt{\ln K/(3U)}$ satisfies

$$\mathbb{E}[R(T)] \leq 2\sqrt{3} \cdot \sqrt{UT \log N},$$

where U is a number known to the algorithm such that

(a) $c_t(e) \leq U$ for all experts e and all rounds t,

(b) $\mathbb{E}[G_t] \leq U$ for all rounds t, where $G_t = \sum_{\text{experts } e} p_t(e)\, c_t^2(e)$.

We will need to distinguish between "experts" in the full-feedback problem and "actions" in the bandit problem. Therefore, we will consistently use "experts" for the former and "actions/arms" for the latter.

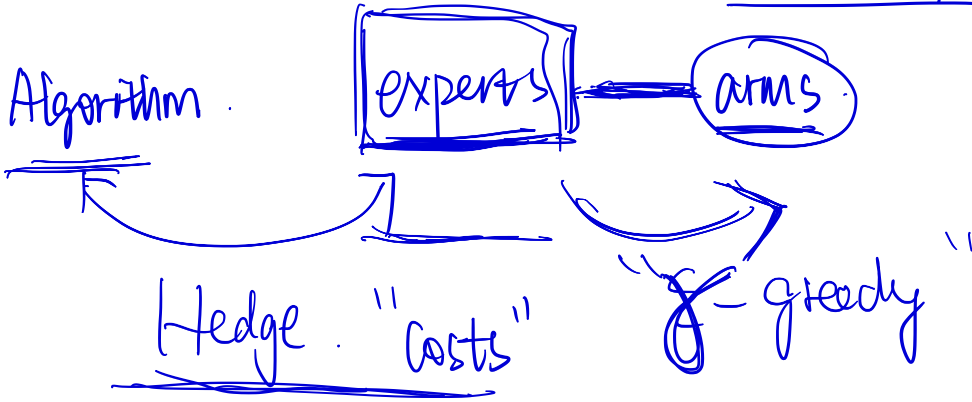
Problem protocol: Adversarial bandits with expert advice

Given: K arms, N experts, T rounds.

In each round $t \in [T]$:

1. adversary picks cost $c_t(a)$ for each arm a,
2. each expert e recommends an arm $a_{t,e}$,
3. algorithm picks arm a_t and receives the corresp. cost $c_t(a_t)$.

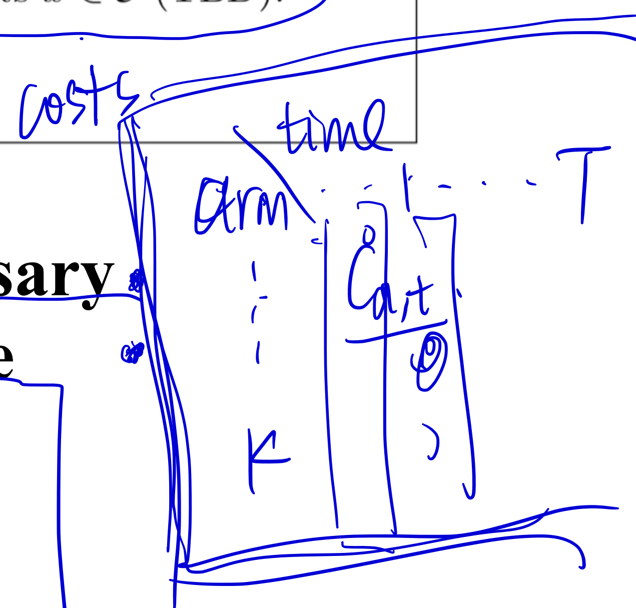
bandit feedback.



Algorithm 6.1: Reduction from bandit feedback to full feedback

Given: set $\mathcal{E}$ of experts, parameter $\epsilon \in (0, \frac{1}{2})$ for `Hedge`.
In each round t,

1. Call `Hedge`, receive the probability distribution p_t over $\mathcal{E}$.
2. Draw an expert e_t independently from p_t.
3. *Selection rule*: use e_t to pick arm a_t (TBD).
4. Observe the cost $c_t(a_t)$ of the chosen arm.
5. Define "fake costs" $\hat{c}_t(e)$ for all experts $x \in \mathcal{E}$ (TBD).
6. Return the "fake costs" to `Hedge`.



*** deterministic oblivious adversary**

*** experts do not learn over time**

recommendation

t

$$
\begin{array}{c|cccc}
 & 1, & 2, & \cdots & T \\
1 & a_{1,1} & \cdots & & a_{T,1} \\
2 & a_{1,2} & \cdots & & a_{T,2} \\
e \quad \vdots & \vdots & & & \\
N & a_{1,N} & \cdots & & a_{T,N}
\end{array}
$$

Unbiased "fake" costs

$\hat{c}$

$$\boxed{\mathbb{E}[\hat{c}_t(e) \mid \vec{p}_t] = c_t(e)} \quad \text{for all experts } e, \tag{6.2}$$

Hedge.

$c_t(a_{t,e})$.

selection according to $\vec{p}_t$.

expert e rec arm $a_{t,e}$ at time t.

$$q_t(a) := \Pr[a_t = a \mid \vec{p}_t] \quad \text{for each arm } a.$$

Using these probabilities, we define the fake costs on each arm as follows:

$$\hat{c}_t(a) = \begin{cases} \frac{c_t(a_t)}{q_t(a_t)} & a_t = a, \\ 0 & \text{otherwise.} \end{cases}$$

The fake cost on each expert e is defined as the fake cost of the arm chosen by this expert: $\hat{c}_t(e) = \hat{c}_t(a_{t,e})$.

Selection rule

w.p. $1-\gamma$. follow e_t

w.p. γ. explore.

Algorithm 6.2: Algorithm `Exp4` for adversarial bandits with experts advice

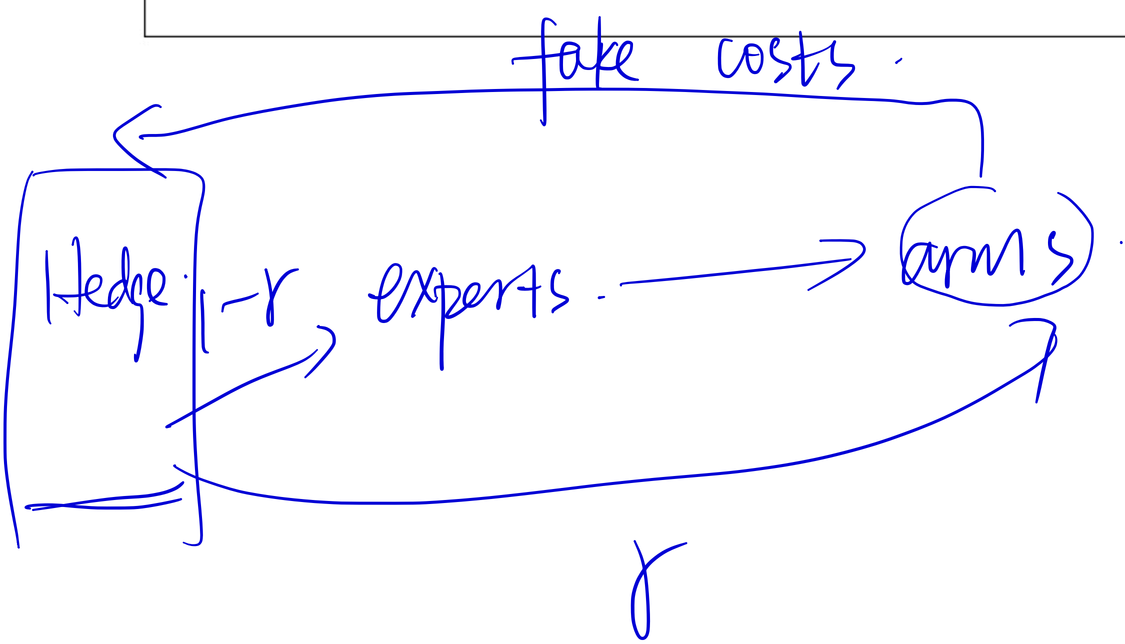
Given: set $\mathcal{E}$ of experts, parameter $\epsilon \in (0, \frac{1}{2})$ for `Hedge`,
exploration parameter $\gamma \in [0, \frac{1}{2})$.

In each round t,

1. Call `Hedge`, receive the probability distribution p_t over $\mathcal{E}$.
2. Draw an expert e_t independently from p_t.
3. *Selection rule*: with probability $1 - \gamma$ follow expert e_t; else pick an arm a_t uniformly at random.
4. Observe the cost $c_t(a_t)$ of the chosen arm.
5. Define fake costs for all experts e:

$$\hat{c}_t(e) = \begin{cases} \frac{c_t(a_t)}{\Pr[a_t = a_{t,e} | \vec{p}_t]} & a_t = a_{t,e}, \\ 0 & \text{otherwise.} \end{cases}$$

6. Return the "fake costs" $\hat{c}(\cdot)$ to `Hedge`.



Improved analysis of EXP4

We obtain a better regret bound by analyzing the quantity

$$\widehat{G}_t := \sum_{e \in \mathcal{E}} p_t(e)\, \hat{c}_t^2(e).$$