Bandits with full feedback

↳ partial feedback (only get reward for chosen action)

Full feedback - observe rewards / costs of all actions (arms), not only chosen one

eg. expert feedback (Online learning with experts)

stock market

Problem setup (full feedback)

K actions / arms, T rounds

At each round t

adversary chooses $c_t(a)$ $\forall a = 1..K$

algorithm picks a_t and incur $c_t(a_t)$

costs of all $\{c_t(a)\}$ reveals

Goal: Do as well as best action.

Online learning with experts

K experts, T rounds

adversary chooses observation $\underset{\sim}{x}_t$ and label $\underset{\sim}{z}_t$

Reveals x_t only, expert prediction $\hat{z}_{i,t}$ $i = 1..K$

Algorithm picks expert $e \in 1, ..., K$

Incurs loss $c_t(e) = c(\underline{\hat{z}_{e,t}}, z_t)$ on seeing true label z_t.

known

Costs for all experts revealed

Note: Known costs for bandits for easy.

Note: Stochastic costs for full feedback is easy.

Play lowest average cost action

$$R(T) \leq \sqrt{\log KT} \left(1 + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \cdots \right) \qquad O\left(\sqrt{\frac{\log KT}{n_t(a)}}\right)$$

$$\frac{1}{\sqrt{t}} \le \frac{2}{\sqrt{t}+\sqrt{t-1}} = 2(\sqrt{t}-\sqrt{t-1})$$

$$\sum \frac{1}{\sqrt{t}} = 2\sqrt{T}$$

$$R(T) = O(\sqrt{T \log KT})$$ vs. Bandit $O(\sqrt{KT \log KT})$

<u>Adversarial setting</u>

Oblivious or Adaptive adversary

costs don't / 'do' depend on algorithm's choices.

(Cumulative) Regret $R(T) = \sum_{t=1}^{T} c_t(a_t) - \sum_{t=1}^{T} c_t(a^*)$

$a^* = \arg\min_a \sum_{t=1}^{T} c_t(a)$ – best in hindsight action.

Pseudo-Regret $\sum_{t=1}^{T} c_t(a_t) - \min_a E[\sum_{t=1}^{T} c_t(a)]$ reduces to iid best in foresight action know distⁿ of costs

For deterministic adversary → Regret

" randomized " → Regret + Pseudo-regret (weaker)

<u>Note</u>: Pseudo-regret cannot exceed regret.

<u>Note</u>: For iid costs/rewards, $\sqrt{T}$ pseudo-regret extend to regret

$\log T$ - - do not . . . .

Binary prediction $\hat{z}_{i,t} \in \{\pm 1\}$

<u>Majority vote</u> of experts who have not made mistake in past.

Thm: If a perfect expert exists, then majority vote makes at most $\log K$ mistakes.

Proof: S_t - experts who make no mistake upto t.

$W_t = |S_t|$, $W_1 = K$

. . . $\in S_t \ \forall t$.

$W_t \geq 1 \quad \forall t$ b/cz of perfect expert ...

If algo make a mistake at t, $W_{t+1} \leq W_t/2$ ←
b/cz majority of experts in S_t are wrong.

Sketch idea:

W_t - notion of total weight on experts

$W_1 \leq \Box$ W_t doesn't increase over time

$W_T \geq \Box$ ← some notion of cost(a^*)

What if no perfect expert?

Weighted Majority Vote

$w_1(a) = 1 \quad \forall a$

For each t
  Make prediction using weighted w_t maj vote.

For each expert i

$w_i \leftarrow w_i$ if correct

$w_i \leftarrow (1-\varepsilon) w_i$ if expert incorrect.

Thm: # mistake by weighted maj vote $\leq \frac{2}{1-\varepsilon}\text{cost}^* + \frac{2}{\varepsilon}\ln K$

$\frac{2}{1-\varepsilon} \downarrow 0$, $\frac{2}{\varepsilon}\ln K \downarrow \ln K$

$\text{cost}^* = \sum_{t=1}^{T} c_t(a^*)$

Beyond binary prediction

Note: Any deterministic algo. has total cost T even for oblivious adversary.

Hedge algorithm $\varepsilon \in (0, \frac{1}{2})$

$w_1(a) = 1 \quad \forall a$

... $p(a) = \frac{w_t(a)}{...}$

→ Sample arm/expert a_t from distribution $p_t = \frac{w_t(a)}{\sum_a w_t(a)}$

observe costs of all arms.

$$w_{t+1}(a) = w_t(a)\,(1-\varepsilon)^{c_t(a)} \qquad \begin{cases} 1. & c_t(a) = 0 \\ 1-\varepsilon & c_t(a) = 1 \end{cases} \quad \text{for binary prediction}$$

<u>Thm</u>: Bounded costs ≤ 1. Consider adaptive adversary s.t. $\sum_{t=1}^{T} c_t(a^*) \leq uT$ for some known u, then Hedge with $\varepsilon = \sqrt{\frac{\ln K}{uT}}$ satisfies.

$$E\Big[\sum_{t=1}^{T} c_t(a_t) - \sum_{t=1}^{T} c_t(a^*)\Big] \leq 2\sqrt{\underbrace{uT}_{\text{cost}(a^*)} \ln K}$$

<u>Proof</u>: ① $W_t = \sum_a w_t(a)$

$W_1 = K$, $W_{T+1} \geq (\text{cost}(a^*))$ ($W_T \geq 1$ for maj vote assuming 1 best expert exists)

$$W_{T+1} > w_{T+1}(a^*) = (1-\varepsilon)^{\sum_t c_t(a^*)} = (1-\varepsilon)^{\text{cost}(a^*)}$$

② $$\frac{W_{t+1}}{W_t} = \sum_a (1-\varepsilon)^{c_t(a)} \frac{w_t(a)}{\sum_a w_t(a)} \quad p_t(a)$$

($W_{t+1} \leq \frac{W_t}{2}$ maj vote for binary prediction)

$$\leq \sum_a (1 - \varepsilon c_t(a))\, p_t(a) = 1 - \varepsilon \sum_a c_t(a) p_t(a)$$

$(1-\varepsilon)^c \leq 1 - \varepsilon c$ if $c \leq 1$ (bounded costs)

$1-\varepsilon \leq (1-\varepsilon c)^{1/c} = 1 - \varepsilon + \ldots$

$$\ln \frac{W_{t+1}}{W_t} < \ln\Big(1 - \varepsilon \sum_a c_t(a) p_t(a)\Big) < -\varepsilon \sum_a c_t(a) p_t(a)$$

$\because 1 - x \leq e^{-x}$

$$\sum_t \ln \frac{W_{t+1}}{W_t} = \ln \prod_t \frac{W_{t+1}}{W_t} = \ln \frac{W_{T+1}}{W_1}$$

$$\sum_t \varepsilon E[c_t(a_t)] = -\sum_t \ln \frac{W_{t+1}}{W_t} = -\ln \frac{W_{T+1}}{W_1} \leq -\ln \frac{(1-\varepsilon)^{\sum_t c_t(a^*)}}{K}$$

$$E\Big[\sum c_t(a_t)\Big] \leq \frac{\ln K}{\varepsilon} + \sum_t c_t(a^*)\, \frac{1}{\varepsilon} \ln \frac{1}{1-\varepsilon}$$

$\sim c_t \qquad \varepsilon \qquad \leq 1+\varepsilon \text{ if } \varepsilon \in [0,\tfrac{1}{2})$

$$E\left[\sum_t c_t(a_t) - \sum_t c_t(a^*)\right] \leq \underbrace{\frac{\ln K}{\varepsilon} + \varepsilon \sum_t c_t(a^*)}$$

$\sum_t c_t(a^*)$, $\ln K$

$$\varepsilon = \sqrt{\frac{\sum_t c_t(a^*)}{\ln K}} \qquad = O\Big(\sqrt{\underbrace{\textstyle\sum_t c_t(a^*)}_{uT} \ln K}\Big)$$

---

① extend to unbounded costs

② Comparator class C_T - action sequence $(a^*, a^*, \dots a^*)$

Regret $\quad \displaystyle\sum_{t=1}^{T} c_t(a_t) - \min_{\substack{y_1 \dots y_T \\ \in C_T}} \sum_{t=1}^{T} c_t(a_1, \dots a_{t-1}, y_t)$ Oblivious adversary

Policy regret $\quad \displaystyle\sum_{t=1}^{T} c_t(a_{1:t}) - \min_{y_1 \dots y_T \in C_T} \sum_{t=1}^{T} c_t(y_{1:t})$ adaptive adversary.

$= \Omega(T)$

times (tallying bandits)