Adversarial bandits: 15/4/20

k-armed adv. bandit:

$(x_t)_{t=1}^{n}$ is a seq. reward vectors with $x_t \in [0,1]^k$;

In each round the learner chooses a distribution over actions $P_t \in \mathcal{P}_{k-1}$.
Then action $A_t \in [k]$ is sampled from P_t, and learner gets x_{tA_t};

Alg:

Adv selects $(x_t)_{t=1}^{n}$ $x_t \in [0,1]^k$

for rounds $t = 1, 2, \dots n$

learner selects $P_t \in \mathcal{P}_{k-1}$ and samples A_t from P_t;

learner sees x_{tA_t}

Policy: $\pi : \left([k] \times [0,1] \right)^* \to \mathcal{P}_{k-1}$, }

mapping history sequences to distributions

over actions; $\underbrace{A_1, X_1, A_2, X_2}\ A_3, \ldots$

$$R_n(\pi, x) = \max_{i \in [k]} \sum_{t=1}^{n} x_{ti} - \mathbb{E}\left[\sum_{t=1}^{n} x_{tA_t} \right]$$

comparing against a fixed action in hindsight;

= Randomness - learner;

- Action choosen in round t may depend upon actions choosen at $s < t$, and the observed rewards.

$$\boxed{R_n^*(\pi) = \sup_{x \in \{0,1\}^{n \times k}} R_n(\pi, x)}$$

If learner is deterministic:

Linear regret can be forced;

Look at the algorithm the learner is using;

If $I_t = 1$, let the rewards be

$$\begin{array}{|c|c|c|c|}\hline 0 & 1 & 1 & \cdots\ 1 \\ \hline\end{array}$$

If $I_t \neq 1$ let rewards be

$$\begin{array}{|c|c|c|c|}\hline 1 & 0 & 0 & \cdots\ 1 \\ \hline\end{array}$$

- Now if $\{t \mid I_t = 1\} \geq n/2$, consider playing arm 2 all the time;

· Regret^2: $\geq n/2$.

- If $\{t \mid I_t = 1\} \leq n/2$,
Consider playing 1 always,
$\therefore \{t \mid I_t \neq 1\} \geq n/2 \therefore$ get a regret of $n/2$ here,

· Forced to look at randomized learners.

Regret is now a random variable,
- we need bounds in high probability or in expectation on R_n.

- Strategies we've seen so far are det.

- Randomization is in the rewards!

So none will work well in the adv. setting

- Converse?

will an adversarial bandit strategy work in the stochastic setting?

- Small expected regret in stochastic setting?

- Let π be an adv bandit strategy and $\nu = (\nu_1, \ldots, \nu_k)$ a stochastic bandit,

$\mathrm{Supp}(\nu_i) \subseteq [0,1] \ \forall i$;

- Let X_{ti} be sampled from ν_i for $i \in [k]$ and $t \in [n]$;

Assume $X_{t,i}$ are mutually indep

$$R_n(\pi, \nu) := \max_{i \in [k]} \mathbb{E}\left[\sum X_{ti} - X_{tA_t}\right]$$

stoch regret.

$$\leq \mathbb{E}\left[\max_{i \in [k]} \sum X_{ti} - X_{tA_t}\right]$$

$$= \mathbb{E}\left[R_n(\pi, X)\right] \leq R_n^*(\pi).$$

Adv regret!

- Worst case stochastic regret is upper bounded by worst case adversarial regret.

- Key idea: A mechanism for estimating rewards of unplayed arms.

Now P_t is the conditional distribution of action played in round t

$$\therefore \quad P_{ti} = \mathbb{P}\left[A_t = i \,\middle|\, A_1, X_1, \ldots, X_{t-1}\right]$$

conditional on history

Define: importance-weighted estimator of x_{ti} is

$$\hat{X}_{ti} = \frac{\mathbb{1}\{A_t = i\} X_t}{P_{ti}}.$$

Let $\mathbb{E}_t[\cdot] = \mathbb{E}\left[\cdot \,\middle|\, A_1, \ldots, X_{t-1}\right]$ conditional exp given history.

If $A_{ti} = \mathbb{1}\{A_t = i\}$ then clearly.

$$X_t A_{ti} = x_{ti} A_{ti}$$

$$\therefore \quad \hat{X}_{ti} = \frac{A_{ti} X_t}{P_{ti}} = \frac{x_{ti} A_{ti}}{P_{ti}}$$

$$\therefore \quad \mathbb{E}_t\left[\hat{X}_{ti}\right] = \mathbb{E}_t\left[\frac{x_{ti} A_{ti}}{P_{ti}}\right]$$

$$= \mathbb{E}_t\left[\frac{x_{ti} A_{ti}}{P_{ti}} \,\middle|\, A_1, X_1, \dots, X_{t-1}\right]$$

$$= \frac{x_{ti}}{P_{ti}} \mathbb{E}_t\left[A_{ti} \mid \dots\right]$$

$$= \frac{x_{ti}}{P_{ti}} \cdot \left[1 \cdot P_{ti} + 0 \cdot (1 - P_{ti})\right]$$

$$= x_{ti}.$$

$\therefore$ $\hat{X}_{ti}$ - an unbiased est. of x_{ti}.

- What about variance?

Now: $V_t[U] = \mathbb{E}_t\left[(U - \mathbb{E}_t(U))^2\right]$

$$\therefore V_t[\hat{X}_{ti}] = \mathbb{E}_t[\hat{X}_{ti}^2] - (\mathbb{E}_t[\hat{X}_{ti}])^2$$

$$= \mathbb{E}_t\left[\frac{A_{ti}\, x_{ti}^2}{P_{ti}^2}\right] - x_{ti}^2$$

$$= \frac{x_{ti}^2}{P_{ti}} - x_{ti}^2 = x_{ti}^2\left(\frac{1-P_{ti}}{P_{ti}}\right)$$

= this can be large if P_{ti} is small and $x_{ti} > \varepsilon$

Another estimator:

$$\hat{X}_{ti} = 1 - \frac{\mathbb{1}\{A_t = i\}(1 - X_t)}{P_{ti}}$$

This too is unbiased;
Using $y_{ti} = 1 - x_{ti}$, $Y_t = 1 - X_t$
& $\hat{Y}_{ti} = 1 - \hat{X}_{ti}$ we get

$$\hat{Y}_{ti} = \frac{\mathbb{1}\{A_t = i\}\, Y_t}{P_{ti}}$$

& again: $\mathbb{E}\left[\hat{Y}_{ti}\right] = 1 - x_{ti}$

Loss based estimator!

$$\mathbb{V}_t\left[\hat{Y}_{ti}\right] = y_{ti}^2\left(\frac{1 - P_{ti}}{P_{ti}}\right)$$

∴ Which is better? $\underbrace{\text{Smaller rewards, for arm } i}$ $\Rightarrow$ $\hat{X}_{ti}$ better

Exp3 algorithm: Exp wts alg for Exp & Exploitation.

Let $\hat{S}_{ti} = \sum_{s=1}^{t} \hat{X}_{si}$, $\hat{X}_{ti} = 1 - \hat{Y}_{ti}$

Clearly $\mathbb{V}_t[\hat{X}_{ti}] = \mathbb{V}_t[\hat{Y}_{ti}]$

Natural to play action with large estimated rewards with bigger probability.

Can map $\hat{s}_{ti} \longrightarrow$ probabilities;

$$\boxed{P_{ti} = \frac{\exp(\eta \hat{s}_{t-1,i})}{\sum_{j=1}^{k} \exp(\eta \hat{s}_{t-1,j})}}$$

η - learning rate;

Use this to get EXP3 algorithm.

1) Input: n, k, η

2) $\hat{S}_{0i} = 0 \quad \forall i$

3) for $t = 1, \dots, n$ do

4) $$P_{ti} = \frac{\exp(\eta \hat{S}_{t-1,i})}{\sum_{j=1}^{k} \exp(\eta \hat{S}_{t-1,j})}$$

5) Sample $A_t \sim P_t$ & observe X_t

6) $$\boxed{\hat{S}_{ti} = \hat{S}_{t-1,i} + 1 - \left(\frac{\mathbb{1}\{A_t = i\}(1 - X_t)}{P_{ti}}\right)}$$

7) endfor;

THM: $x \in [0,1]^{n \times k}$; set $\eta = \sqrt{\log(k)/(nk)}$.

Then $R_n(\pi, x) \leq 2\sqrt{nk \log k}$.