April 22, 2020:

Adversarial bandits: High probability bound (from Bubeck & Bianchi)

- Issue - the variance of loss $\tilde{Y}_{i,t}$ is $O(1/p_{i,t})$.

- One idea: Ensure that $p_{i,t}$ is not too small, by playing exp-3 policy with prob $(1-\varepsilon)$ & uniform with prob ε, so that each arm is played with $p_{i,t} > \varepsilon/K$.

=

Want a regret of $\sqrt{n}$; since in each round you may collect constant regret, ε can't be too large.

In fact $\varepsilon \le \frac{1}{\sqrt{n}}$, so that over

n rounds the uniform dist will contribute at most $O(\sqrt{n})$ regret!

But then the variance of cumulative regret can be $\sqrt{n}$ per round & so

$\sqrt{n} \cdot n = n^{3/2}$; $\underbrace{\text{So this doesn't work;}}$

= We work with gain $- g_{i,t} = \underline{1 - l_{i,t}}$;

= Introduce a $\underline{\text{bias}}$ in the gain estimate.

Lemma: Let $\beta \le 1$ & Set

$$\boxed{\tilde{g}_{i,t} = \frac{g_{i,t} \mathbb{1}_{I_t = i} + \beta}{p_{i,t}}}$$

Then with prob $1-\delta$,

$$\sum_{t=1}^{n} g_{i,t} \le \sum_{t=1}^{n} \tilde{g}_{i,t} + \frac{\ln(1/\delta)}{\beta}.$$

<u>Proof:</u> $\mathbb{E}_t$ - be conditional exp, conditional on $I_1, I_2, \ldots, I_{t-1}$

$$\mathbb{E}_t \exp\left(\beta g_{it} - \frac{\beta\left(g_{it}\mathbb{1}_{I_t=i}+\beta\right)}{p_{it}}\right)$$

$$= \mathbb{E}_t \exp\left(\beta g_{it} - \frac{\beta g_{it}\mathbb{1}_{I_t=i}}{p_{it}} - \frac{\beta^2}{p_{it}}\right)$$

$$= \mathbb{E}_t\left[\exp\left(\beta g_{it} - \frac{\beta g_{it}\mathbb{1}_{I_t=i}}{p_{it}}\right)\exp\left(\frac{-\beta^2}{p_{it}}\right)\right]$$

$$= \exp\left(-\beta^2/p_{it}\right)\mathbb{E}_t\left[\exp\left(\underbrace{\beta g_{it} - \frac{\beta g_{it}\mathbb{1}_{I_t=i}}{p_{it}}}_{X}\right)\right]$$

$$\exp(x) \le 1 + x + x^2 \quad \forall x \le 1.$$

$$\le \exp\left(\frac{-\beta^2}{p_{it}}\right)\left(1 + \mathbb{E}_t(x) + \mathbb{E}_t(x^2)\right)$$

$$\le \exp\left(\frac{-\beta^2}{p_{it}}\right)\left(1 + 0 + \frac{\beta^2 g_{it}^2}{p_{it}} - \underset{\uparrow \atop \text{+ve}}{(\ )}\right)$$

$$\le \exp\left(\frac{-\beta^2}{p_{it}}\right)\underbrace{\left(1 + \frac{\beta^2 g_{it}^2}{p_{it}}\right)}_{\le e\left(\frac{\beta^2 g_{it}^2}{p_{it}}\right)}$$

$$\le 1.$$

$$\therefore \mathbb{E}\left[\exp\left(\overbrace{\beta\sum_{t=1}^{n} g_{it} - \beta\sum_{t=1}^{n}\left(\frac{g_{it}\mathbb{1}_{I_t = i} + \beta}{p_{it}}\right)}^{X}\right)\right]$$

$$\le 1.$$

$$\Pr\left[X > \ln(\delta^{-1})\right] = \Pr\left[e^X > 1/\delta\right]$$

$$\le \delta\, \mathbb{E}\left[e^X\right] \le \delta$$

$\lambda.\ \Pr\left[\beta \sum_{t=1}^{n} g_{it} - \beta \sum_{t=1}^{n} \frac{g_{it} \mathbb{1}_{I_t=i} + \beta}{p_{i,t}} \le \ln(1/\delta)\right]$

$\geq 1-\delta;$

Exp 3.P:

Input: $\eta, \gamma, \beta \in (0,1)$

$p_1 \triangleq$ uniform dist on $1, \dots, K$

for $t=1, \dots n$

① Draw arm I_t from prob dist p_t

(2) Compute estimated gain,

$$\tilde{g}_{i,t} = \frac{g_{it} \mathbb{1}_{I_t=i} + \beta}{p_{i,t}}$$

• Update estimated cum gain

$$\tilde{G}_{i,t} = \sum_{s=1}^{t} \tilde{g}_{i,s}$$

(3) $p_{t+1} = (p_{t+1,1}, \dots, p_{t+1,K})$

$$p_{i,t+1} = \frac{(1-\gamma)\exp(\eta \tilde{G}_{i,t})}{\sum_{k=1}^{K} \exp(\eta \tilde{G}_{k,t})} + \frac{\gamma}{K}$$

end for:

THM: Set $\beta = \sqrt{\frac{\ln(K\delta^{-1})}{nK}}$; $\gamma = 1.05\sqrt{\frac{K\ln K}{n}}$

$\eta = 0.95\sqrt{\frac{\ln K}{K}}$;

Get:

with prob $1-\delta$,

$$R_n \leq 5.15 \sqrt{nK \ln(K/\delta)}$$

= Contextual bandits:

- learner has access to extra info.

ex: movie recommendation:

- we should look at contextual info, past history of movies, and also the content / type of movie when making a recommendation!

- Need to devise algos which use this contextual info.

- <u>Basic example:</u>

Bandits with side info;

A fixed set of contexts $\mathcal{C}$;

rounds are marked by contexts $c_1, \ldots, \in \mathcal{C}$;

Learner must learn a mapping

$$g : \mathcal{C} \longrightarrow \{1, \ldots, K\}.$$

<u>Idea:</u> Run a different EXP3 on each context!

=

-

$c = |\mathcal{E}|$;

Run one exp3 on each context.

n_c = # times context c is played;

$$R_n = \mathbb{E}\left[\sum_{c \in \mathcal{E}} \max_{k \in K} \left(\sum_{t : c_t = c} l_{I_t,t} - l_{k,t}\right)\right]$$

$$= \sum_{c \in \mathcal{E}} \max_{k=1 \dots K} \sum_{t : c_t = c} \left(l_{I_t,t} - l_{k,t}\right)$$

$$\leq \sum_{c \in \mathcal{E}} \sqrt{2 n_c K \ln K}$$

concave $\sqrt{\ }$

$$= |\mathcal{E}| \sum_{c \in \mathcal{E}} \frac{1}{|\mathcal{E}|} \sqrt{2 n_c K \ln K}$$

$\mathbb{E}\left(g(X)\right)$, $= \sqrt{X}$, $\leq g\left(\mathbb{E}(X)\right)$.

$$\leq |\mathcal{E}| \sqrt{\frac{\sum_c 2 n_c K \ln K}{|\mathcal{E}|}}$$

$$= |\mathcal{E}| \sqrt{\frac{2 n K \ln K}{|\mathcal{E}|}} = \sqrt{2 n K |\mathcal{E}| \ln K}.$$

- Playing with experts:

when $|\mathcal{C}|$ is large - bad idea!

- Users with similar demographics like similar movies!

$\therefore$ Contexts are structured!

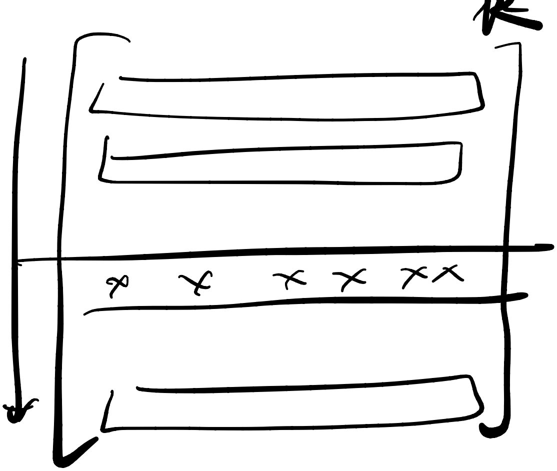
$$\Phi = \{ f : \mathcal{C} \to k \}$$

set
$$R_n = \mathbb{E}\left[\max_{\phi \in \Phi} \sum_{t=1}^{n} \left(x_{t\phi(t)} - X_t \right) \right]$$

- At the beginning of each round experts announce their predictions!

- In fact experts give a prob dist over actions; (experts are randomized)

- The expert advice in round t

$E^t_j =$ 

R_n – measured w.r.t best expert in hindsight;

$$R_n = \mathbb{E}\left[\max_{m\in M}\sum_{t=1}^{n} E^t_m x_t - \sum X_t\right]$$

$x_t = (x_{t1}, \ldots, x_{tk})$

Exp 4:

Expert; Exp wts, Explore, Exploit

Input: n, k, M, η, γ

2) $Q_1 = (1/M, \ldots, 1/M)$

3) for $t = 1, \ldots n$

4) Receive advice E^t

5) Choose $A_t \sim P_t$, $P_t = Q_t E^t$

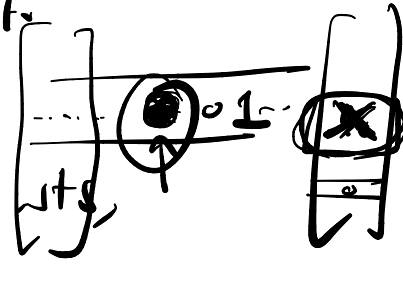
6) Receive reward $X_t = x_{tA_t}$

7) Estimate rewards

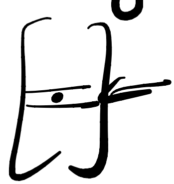
$$\boxed{\hat{X}_{ti} = 1 - \frac{\mathbb{1}\{A_t = i\}(1 - X_t)}{P_{ti} + \gamma}}$$

8) Propagate rewards to experts

$$\tilde{X}_t = E^t \hat{X}_t$$



9) Update Q_t using exp wts



$$\theta_{t+1,i} = \frac{\exp(\eta \tilde{x}_{ti}) \theta_{ti}}{\sum_{j=1}^{m} \exp(\eta \tilde{x}_{tj}) \theta_{tj}}$$

10 hd;

___ x ___ x ___ x ___ x ___

$\sqrt{nK \ln K |\mathcal{E}|}$.

$\log |\mathcal{E}|$.

___ x ___