- Contextual bandits:

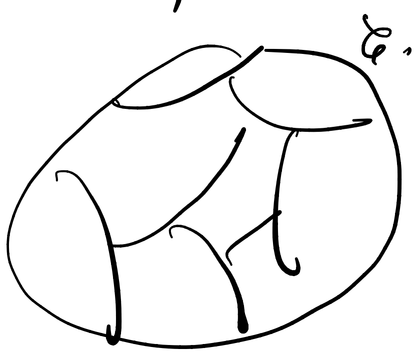
Seeking to bound

$$\Phi = \{ \varphi \mid \varphi : \underbrace{\mathcal{C}}_{\text{contexts}} \to [\underbrace{k}_{\#\text{arms}}] \}$$

$$\underbrace{R_n = \mathbb{E}\left[\max_{\phi \in \Phi} \sum_{t=1}^{n} x_{t\,\phi(c_t)} - X_t \right]}_{\text{pseudo regret;}}$$

- Instead of playing with all functions from $\mathcal{C} \to k$ we could restrict to a smaller subset of functions.
  Regret will $\downarrow$, or reward $\uparrow$;

Examples:



Partition $\mathcal{C}$;

Play the same arm on all contexts in a given part;

Regret will depend upon # parts;

(2) Similarity function:

$$s : \mathcal{C} \times \mathcal{C} \to [0,1]$$

$$\Phi = \left\{ \varphi : \mathcal{C} \to \mathcal{A} \;\middle|\; \frac{1}{|\mathcal{C}|^2} \sum_{c,d \in \mathcal{C}} (1 - s(c,d)) \, \mathbb{1}[\varphi(c) \neq \varphi(d)] \leq \theta \in [0,1] \right\}$$

- Dissimilarity function is bounded;
- Play against functions which will send similar contexts to the same arm.

- If similarity b/w contexts is low, then we will pull different arms for these contexts; But the # functions $\phi \in \Phi$ is also not too many, $\because$ $1 - s(c,d)$ is large, $\therefore$ our regret is better!

- Pick a collection of predictors, $\phi_1, \dots, \phi_M$; each $\phi_i : \mathcal{C} \to [k]$;

We can use a bandit algorithm to compete with the best of these in an online fashion;

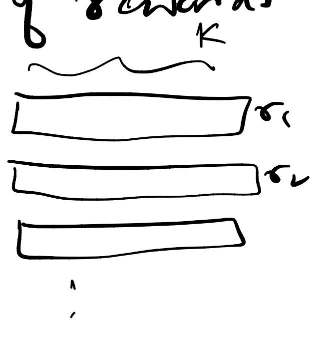
Adv: ϕ_i's can be obtained offline using batch training.

- So natural to analyze contextual bandits in the framework of experts;

VER (1): simplest; Follow the expert.

k experts; sequence of rewards $r_1, r_2, \dots$

$r_i \in [0,1]^K$;



- Every day choose a distribution p_t over experts;
- At the end of the day see the quality of EVERY expert;

$$\text{Regret}(T) := \max_i \sum_{t=1}^{T} r_{t,i} - \sum_{t=1}^{T} \langle p_t, r_t \rangle$$

or $$\text{Regret}(T) := \max_p \sum_{t=1}^{T} \langle p, r_t \rangle - \sum_{t=1}^{T} \langle p_t, r_t \rangle$$ 2

Best convex combination of experts;

Want a bound on max regret;

Start with a wt of 1 for each expert;
Given rewards, r^t

$$w_i^{t+1} = w_i^t (1 + \varepsilon r_i^t)$$

If r_i^t is large ↑ the wt of this expert;

- Given wts set $\boxed{\phi^t = \sum_i w_i^t}$

· Let $p^t = \dfrac{w^t}{\phi^t}$, $w^t = \langle w^t_1, \ldots, w^t_k \rangle$.

- we show that if the horizon is $> \dfrac{\ln k}{\varepsilon^2}$

then the regret ② satisfies,

horizon;

$$② \leq \frac{\ln k}{\varepsilon} + \varepsilon n.$$

$$\underbrace{\frac{②}{n}}_{\text{average regret over the horizon}} \leq \frac{\ln(k)}{n\varepsilon} + \varepsilon$$

$$\leq \varepsilon + \frac{\ln(k)}{n\varepsilon}$$

Set: $n \geq \dfrac{\ln(k)}{\varepsilon^2}$

AVG regret $\leq \varepsilon + \varepsilon = 2\varepsilon$.

$$\phi^{t+1} = \sum w_i^{t+1} = \sum w_i^t (1+\varepsilon r_i^t)$$

$$= \sum p_i^t \phi^t (1+\varepsilon r_i^t)$$

$$\phi^{t+1} = \phi^t + \varepsilon \phi^t \langle p^t, r^t \rangle$$

$$= \phi^t \left[1 + \varepsilon \langle p^t, r^t \rangle \right]$$

$$\phi^{t+1} \le \phi^t \cdot e^{\varepsilon \langle p^t, r^t \rangle}$$

$$\le \phi^0 \cdot e^{\varepsilon \sum \langle p^t_j, r^t \rangle}$$

$$= k \cdot e^{\varepsilon \sum_{t=1}^{n} \langle p^t, r^t \rangle}$$

Now $\phi^{t+1} \ge w_i^{t+1} \quad \forall i,$

$$= \prod_{t=1}^{n} (1+\varepsilon r_i^t);$$

<u>choose</u>

$\varepsilon \le 1/2$

$\because r_i^t \le 1$

$\le 1/2.$

Now for $|x| \le 1/2$,

$$\ln(1+x) \ge x - x^2$$

$$\phi^{t+1} \ge \prod_{t=1}^{n} e^{\varepsilon r_i^t - \varepsilon^2 (r_i^t)^2} = e^{\varepsilon \sum r_i^t} \cdot e^{-\varepsilon^2 \sum (r_i^t)^2}$$

$$k \cdot e^{\varepsilon \sum_{t=1}^{n} \langle p^t, r^t \rangle} \geq e^{\varepsilon \sum_{t=1}^{n} r_i^t} \cdot e^{-\varepsilon^2 \sum_{t=1}^{n} (r_i^t)^2}$$

$$\ln k + \varepsilon \sum_{t=1}^{n} \langle p^t, r^t \rangle \geq \underbrace{\varepsilon \sum r_i^t - \varepsilon^2 \sum (r_i^t)^2}$$

$$\frac{\ln(k)}{\varepsilon} + \sum_{t=1}^{n} \langle p^t, r^t \rangle \geq \sum_{t=1}^{n} r_i^t - \varepsilon \sum_{t=1}^{n} (r_i^t)^2$$

or:
$\forall i$

$$\sum r_i^t - \sum_{t=1}^{n} \langle p^t, r^t \rangle \leq \frac{\ln(k)}{\varepsilon} + \varepsilon \sum_{t=1}^{n} (r_i^t)^2$$

$[-1,1]$

$$\sum_{t=1}^{n} r_i^t - \sum_{t=1}^{n} \langle p^t, r^t \rangle \leq \frac{\ln(k)}{\varepsilon} + \varepsilon n$$

Taking a convex comb of these inequalities

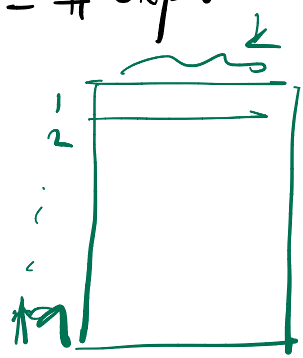
$$\boxed{\max_p \sum_{t=1}^{n} \langle p, r^t \rangle - \sum_{t=1}^{n} \langle p^t, r^t \rangle \leq \frac{\ln(k)}{\varepsilon} + \varepsilon n}$$

$$\text{Avg} \leq \frac{\ln(k)}{n\varepsilon} + \varepsilon$$

QED

Regret analysis with exp ntr:

$M = \#$ expert.



1) Input n, k, M, η, γ
2) $Q_1 = (1/M, \ldots, 1/M)$
3) for $t = 1$ to n
4) Get advice $E^{(t)}$
5) Choose action $A_t \sim P_t$, $P_t = Q_t E^{(t)}$
6) Get $X_t = x_{tA_t}$
7) Set $\hat{X}_{ti} = 1 - \frac{\mathbb{1}(A_t = i)(1 - X_t)}{P_{ti} + \gamma}$
8) Propagate rewards $\boxed{\tilde{X}_{ti} = E^{(t)\hat{}} X_{ti}}$
9) $$Q_{t+1,i} = \frac{\exp(\eta \tilde{X}_{ti}) Q_{ti}}{\sum_j \exp(\eta \tilde{X}_{tj}) Q_{tj}} \quad \forall\, i \in [M]$$
10) end for

• Analyze when $\gamma = 0$:

Thm: $\gamma = 0$; $\eta = \sqrt{2\log(M)/nk}$ & let

R_n = regret after n rounds; Then

$$R_n \le \sqrt{2nk\log(M)} \le \sqrt{nk\log k\,|C|}$$

Recall we showed: $\forall i$

$$\hat{S}_{ni} - \hat{S}_n \le \frac{\log(k)}{\eta} + \eta \sum_{t=1}^{n} \sum_{j=1}^{k} P_{tj}\, \hat{X}_{tj}^2$$

strengthen to $\frac{\eta}{2} \sum_{t=1}^{n} \sum_{j=1}^{k} P_{tj}\, \hat{Y}_{tj}$

That proof yields:

$$\sum_{t=1}^{n} \tilde{X}_{t,m^*} - \sum_{t=1}^{n} \sum_{m=1}^{M} Q_{tm} \tilde{X}_{tm} \le \frac{\log(M)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \sum_{m=1}^{M} P_{tm} (1 - \hat{X}_{tm})^2.$$

$E^{(1)}, A_1, \ldots, A_{t-1}, E^{(t)}$ be the history,

$$\mathbb{E}_t[\quad] = \mathbb{E}[\quad \mid \text{history}]$$

Let $\boxed{m^* = \underset{m \in [M]}{\operatorname{argmax}} \sum_{t=1}^{n} E_m^\top x_t}$ ← expected return of expert m.

<u>Now:</u> $\mathbb{E} = \mathbb{E}[\mathbb{E}_t]$

$$\sum_{t=1}^{n} \underbrace{\tilde{X}_{t,m^*}}_{} - \sum_{t=1}^{n} \sum_{m=1}^{M} P_{tm} \tilde{X}_{tm} \le \frac{\log(M)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \sum_{m=1}^{M} \underbrace{P_{tm}}_{} \left(1 - \hat{X}_{tm}\right)^2 \quad **$$

- $\hat{X}_{ti}$ — unbiased $(\gamma = 0)$ ∴ $\mathbb{E}_t[\hat{X}_t] = x_t$

$$\underbrace{\mathbb{E}_t[\tilde{X}_t]}_{= E^{(t)} x_t} = \mathbb{E}_t[E^{(t)} \hat{X}_t] = E^{(t)} \underbrace{\mathbb{E}_t[\hat{X}_t]}_{x_t}$$

Now

$$\boxed{\mathbb{E}[\tilde{x}] = \mathbb{E}\left[\mathbb{E}_t[\tilde{x}]\right]}$$

Applying $\mathbb{E}$ on both sides and using the above, (first apply $\mathbb{E}_t$ & then $\mathbb{E}$ on both sides)

$$\mathbb{E}\left[E^{(t)}x_t - \sum_{t=1}^{n} x_t\right] = R_n \leq \frac{\log M}{\eta} + \frac{\eta}{2}\sum_{t=1}^{n}\sum_{m=1}^{M}\mathbb{E}\left[q_{tm}(1-\tilde{x}_{tm})^2\right]$$

$$\therefore R_n \leq \frac{\log M}{\eta} + \frac{\eta}{2}\sum_{t=1}^{n}\sum_{m=1}^{M}\mathbb{E}\left[q_{tm}(1-\tilde{x}_{tm})^2\right]$$

Work with $\hat{Y}_{ti} = 1-\hat{x}_{ti}$; $y_{ti} = 1-x_{ti}$

$\tilde{Y}_{tm} = 1-\tilde{x}_{tm}$

Now: $\tilde{Y}_t = E^{(t)}\hat{Y}_t$.

Use notation $A_{ti} \triangleq \mathbb{1}\{A_t = i\}$

Then $\hat{Y}_{ti} = \dfrac{A_{ti}\, y_{ti}}{P_{ti}}$

$$\mathbb{E}_t\left[\tilde{Y}_{tm}^2\right] = \mathbb{E}_t\left[\left(\frac{E^{(t)}_{mA_t}\, y_{tA_t}}{P_{tA_t}}\right)^2\right]$$

$$= \sum_{i=1}^{k} \frac{\left(E^{(t)}_{mi}\, y_{ti}\right)^2}{P_{ti}} \leq \sum_{i=1}^{k} \frac{E^{(t)}_{mi}}{P_{ti}}.$$

$$\therefore \mathbb{E}\left[\sum_{m=1}^{M} Q_{tm}\left(1-\tilde{X}_{tm}\right)^2\right]$$

$$= \mathbb{E}\left[\mathbb{E}_t\left[\sum_{m=1}^{m} Q_{tm}\left(\tilde{Y}_{tm}^2\right)\right]\right]$$

$$\leq \mathbb{E}\left[\sum_{m=1}^{m} Q_{tm} \sum_{i=1}^{k} \frac{E^{(t)}_{mi}}{P_{ti}}\right]$$

$$= \mathbb{E}\left[\sum_{i=1}^{k} \underbrace{\frac{\sum_{m=1}^{m} Q_{tm} E^{(t)}_{mi}}{P_{ti}}}_{1}\right] = \underline{k}. \qquad \because P_t = Q_t E^{(t)}$$

$$\eta = \sqrt{\frac{2\log M}{nk}}.$$

$$\boxed{\therefore R_n \leq \frac{\log M}{\eta} + \frac{\eta\, n k}{2}} = \sqrt{2nk\log M}$$