HW3. Pb.2. Piazza @29.

Online learning w/ bandit feedback:

§: stochastic multi-armed bandits.

Recall product recommendation:

For $t = 1, 2, \dots T$:

environment draws $\ell_t \in [0,1]^K$ (not revealed to learner)

learner selects action (arm) $a_t \in \{1 - K\}$.

learner suffers loss $\ell_t(a_t)$.

& sees.

and it does not see $\{\ell_t(a) : a \neq a_t\}$.

$\ell_t^{(i)}$:

$\in [0,1]^K$

$(\ell(a) : a \in \{1-K\})$

$a \in \mathbb{R}^d$.

Assume: $\forall a$: $(\ell_t(a))_{t=1}^T$ are drawn iid

$\ell(a) \in [0,1]$ from a distn over $[0,1]$. with ...

Goal: minimize: $\sum_{t=1}^{T} \ell_t(a_t)$

$T \cdot \ell(a)$.

t \ a	1	--	K
1			
:		$\ell_t(a)$	
T			

Ex: slot machine game

challenge: ① learn which machine is rewarding → explore

② play machines that are rewarding → exploitation

performance metric:

pseudo-regret:

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t)\right] - T \cdot \ell(a^*)$$

$$\mathbb{E}\left[\min_a \sum_{t=1}^{T} \ell_t(a)\right]$$

$$a^* = \underset{a \in \{1,\dots,K\}}{\operatorname{argmin}} \ell(a).$$

$$\mathbb{E}_{\ell_t \sim D}[\ell_t(a_t)] = \ell(a_t)$$

$$\Rightarrow \mathbb{E}[\ell_t(a_t)] = \mathbb{E}[\ell(a_t)].$$

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T}\left(\ell(a_t) - \ell(a^*)\right)\right]$$

ε-greedy: w.p. ε $a_t \sim \text{unif}(\{1..K\})$

w.p. $1-\varepsilon$. $a_t = \arg\min_a \bar{\ell}_{t-1}(a)$

$$\bar{\ell}_t(a) = \begin{cases} \dfrac{\sum_{s=1}^{t} I(a_s=a)\,\ell_s(a_s)}{m_t(a)}, & m_t(a) > 0 \\ 0 \text{ or } 1, & m_t(a) = 0 \end{cases}$$

$m_t(a) = \sum_{s=1}^{t} I(a_s = a)$: # trials of a up to t

w/ appropriate tuning of ε. $R_T \le O(K^{\frac{1}{3}} T^{\frac{2}{3}})$.

Explore-then-commit:

explore: For $t = 1, 2 \cdots M$: $\approx O(M)$.

(take action $1 \cdots K$ in round robin)

$a_t = ((t-1) \bmod K) + 1$.

t	a_t
1	1
2	2
⋮	⋮
K	K
$K+1$	1
⋮	⋮

$\frac{M}{K}$ times

$\hat{a} = \arg\max_{a \in \{1..K\}} \bar{\ell}_M(a)$. $\in [\ell(a) \pm \sqrt{\frac{K}{M}}]$.

exploitation: For $t = M+1 \cdots T$:

$a_t = \hat{a}$ $\to \sqrt{\frac{K}{M}}$. $\approx O((T-M)\sqrt{\frac{K}{M}})$? ? ?

roughly speaking $R_T \le O\left(M + T\sqrt{\frac{K}{M}}\right)$.

$= O(K^{\frac{1}{3}} T^{\frac{2}{3}})$

$M = K^{\frac{1}{3}} T^{\frac{2}{3}} \Rightarrow R_T \le O(K^{\frac{1}{3}} T^{\frac{2}{3}})$.

Analysis:

Lemma: ($K \le T$), For any online bandit learner, there exists an event E. "clean event",

$P(E) \ge 1 - \frac{2}{T}$, $\ell \in [0,1]$ $\ell \in [0,1]$

$$\forall a, \forall t, \quad \left| \bar{\ell}_t(a) - \ell(a) \right| \le 2\sqrt{\frac{\ln T}{m_t(a)+1}}.$$

Handle corner case $m_t(a)=0$.

Pf: uses Hoeffding ineq & union bound

but is not trivial.

(e.g. Slivkins: Introduction to MAB).

Regret analysis:

$1 = I(E^c) + I(E)$

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \left(\ell(a_t) - \ell(a^*) \right) \right]$$

$$\le T \cdot P(E^c) + \mathbb{E}\left[\left(\sum_{t=1}^{T} \ell(a_t) - \ell(a^*) \right) I(E) \right]$$

$P(E^c) \le 2$.

$$\sum_{t=1}^{M} l(a_t) - l(a^*) + \sum_{t=M+1} l(a_t) - l(a^*)$$

$$\leq M. \qquad \leq (T-M) \cdot 4 \sqrt{\frac{\ln T}{\frac{M}{K}+1}}$$

$$\leq 2 + M + 4T \cdot \sqrt{\frac{\ln T \cdot K}{M}}.$$

$$M = T^{2/3} K^{1/3}$$

$$= \tilde{O}(K^{\frac{1}{3}} T^{\frac{2}{3}}).$$

Can we do better?

Idea: using confidence bounds to guide exploration

UCB algorithm (Auer, et al, 2002).

upper confidence bound

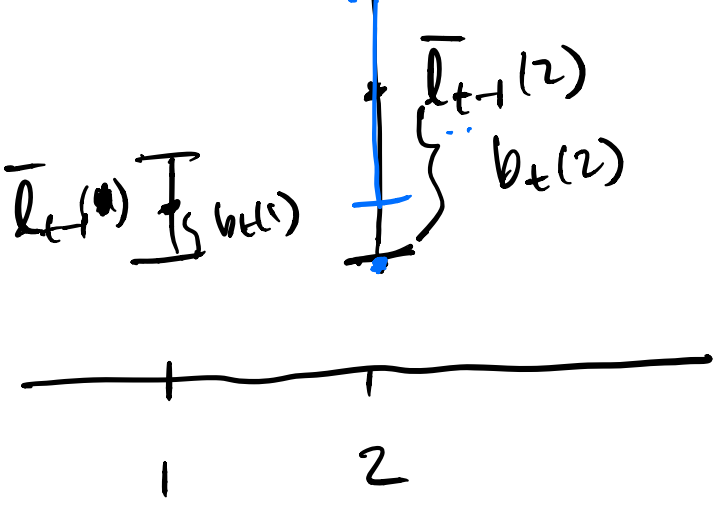
For $t = 1, 2, \cdots T$:

choose $a_t = \underset{a \in \{1 \cdots K\}}{\operatorname{argmin}} LCB_t(a)$.

empirical loss — uncertainty of a

$$LCB_t(a) = \hat{l}_{t-1}(a) - b_t(a),$$

$$b_t(a) = 2 \cdot \sqrt{\frac{\ln T}{m_{t-1}(a) + 1}}.$$



Analysis:

Def. $\Delta(a) = l(a) - l(a^*)$. is the suboptimality gap of action a.

$\Delta(a)$ is large $\Rightarrow$ a is "easy"

$\Delta(a)$ is small $\Rightarrow$

Thm: UCB satisfies:

① $R_T \le \sum_{a:\Delta(a)>0} \frac{16 \ln T}{\Delta(a)} + 3K$ (gap dependent)

say $\Delta(a) = \frac{1}{T}$. $\forall a \neq a^*$.

② $R_T \le \mathcal{O}(\sqrt{TK})$. (gap independent)

Matching lower bound:

Thm. $\exists$ const. $C > 0$ for any algorithm A, there exists a stochastic MAB environment, such that A satisfies $R_T \geq C \cdot \sqrt{TK}$.

Pf of upper bounds:

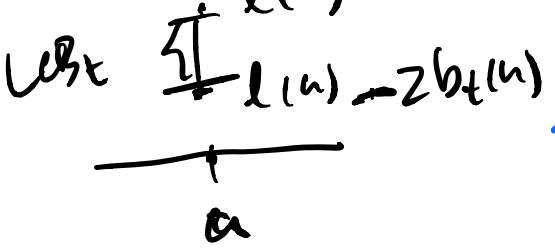
Lemma 1: on event E:

$\forall a, \forall t$,

(a): $LCB_t(a) \leq \ell(a)$ (honest)

(b) $LCB_t(a) \geq \ell(a) - 2b_t(a)$.

(tightness).



$\geq l_{t-1}(a) - b_t(a)$

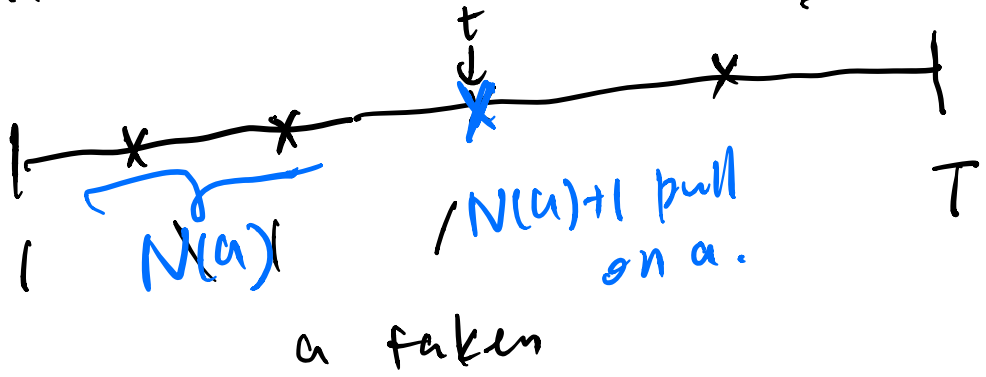
$\geq \ell(a) - b_t(a) - b_t(a)$

Lemma 2: $\forall a. \ \mathbb{E}[m_T(a)] \leq \frac{16 \ln T}{\Delta(a)^2} + 3$.

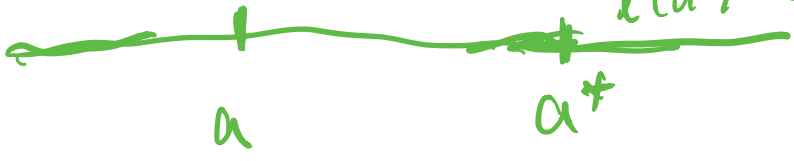
Pf: key claim: on E,

$$m_T(a) \leq \left\lceil \frac{16 \ln T}{\Delta(a)^2} \right\rceil = N(a).$$

If $m_T(a) \geq N(a)+1$, then there must exist some t. $a_t = a$, $m_{t+1}(a) = N(a)$.



$LCB_t(a) \leftarrow \left[\cdot\ \ell(a) - 2b_t(a),\ \cdot\ \ell(a^*) \right] \rightarrow LCB_t(a^*)$



Q: $m_{t-1}(a) = N(a) \Rightarrow 2b_t(a) < \Delta(a)$.

(exercise).

Lemma 1 (b)

$$LCB_t(a) \geq \ell(a) - 2b_t(a)$$

$$> \ell(a) - \Delta(a)$$

$$= \ell(a^*)$$

Lemma 1 (a).

$\geq$ $LCB_t(a)$

this contradicts with $a_t = a$.

$$\Rightarrow \mathbb{E}[m_T(a)] \qquad \leq T$$

$$= \mathbb{E}[m_T(a)\, I(E)] + \mathbb{E}[m_T(a) I(E^c)]$$

$$\leq T \cdot P(E^c)$$

$$\leq \left\lceil \frac{16 \ln T}{\Delta(a)^2} \right\rceil + 2$$

$\lceil x \rceil \leq x + 1$.

$$\leq \frac{16 \ln T}{\Delta(a)^2} + 3.$$

Concluding the regret analysis:

gap dependent bound:

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} l(a_t) - l(a^*) \right]$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} \Delta(a_t)\right]$$

$$= \mathbb{E}\left[\sum_{a=1}^{K} \sum_{t:\, a_t = a} \Delta(a)\right]$$

$$m_T(a) \cdot \Delta(a)$$

$$= \sum_{a=1}^{K} \Delta(a) \cdot \mathbb{E}\left[m_T(a)\right].$$

$$= \sum_{a:\Delta(a)>0} \Delta(a)\, \mathbb{E}\left[m_T(a)\right]$$

Lemma 2

$$\leq \sum_{a:\Delta(a)>0} \Delta(a) \cdot \left(\frac{16 \ln T}{\Delta(a)^2} + 3\right)$$

$$= \sum_{a:\Delta(a)>0} \frac{16 \ln T}{\Delta(a)} + 3K.$$

Next time:

- gap independent bound for UCB.
- adversarial MAB problem.