Lec 9

Page 1

CS59009 - RL

Agenda:

- MOSS & Bernoulli bandit, Adverserial bandit
- contextual and linear bandit
- Linear regression.

Minimax Optimal Strategy is Stochastic multi-armed bandit (MOSS)

With a different analysis, we can set

MOSS $\rightarrow$

$$UCB_i(\delta,t) = \sqrt{\frac{4}{T_i(t-1)} \log^+\left(\frac{n}{K T_i(t-1)}\right)} + \hat{\mu}_{i,t-1}$$

where $\log^+(x) = \log(\max\{1,x\})$

Theorem. For any 1-sub-Gaussian K-armed bandit, the regret of MOSS satisfies

$$R_n \leq 38\sqrt{Kn} + \sum_{i=1}^{K} \Delta_i$$

proof: Chapter 9. Bandit Algorithm.

You are given some ~~a~~ side information.

Bernoulli bandit:

Consider a K-armed bandit with $\mu_i \in [0,1]$ for all i and $X_t \sim B(\mu_{A_t})$

Def: Relative entropy between Bernoulli distributions with parameters $p, q \in [0,1]$ is

$$d(p,q) = p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right)$$

Relative entropy is also known as Kullback-Leibler (KL) divergence.

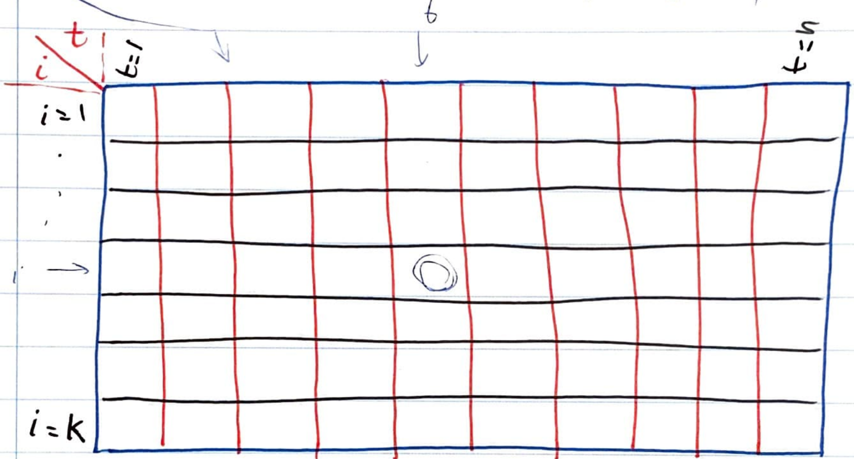
KL-UCB

$$A_t = \operatorname*{argmax}_i \max\left\{\mu \in [0,1] : d(\hat{\mu}_{i,t-1}, \mu) < \frac{\log(1+t\log^2 t)}{T_i(t-1)}\right\}$$

Chapter 10 Bandit Algorithm.

Adversarial bandit:

- A K-armed adversarial bandit is an arbitrary sequence of reward vectors $\nu = (x_1 \dots x_n)$ where $x_t \in [0,1]^K$ for each $t \in [n]$



$\hookrightarrow$ (i,t) entry represent the rewart at time t if arm i is pulled.

In each round ~~of~~ the learner chooses an action $A_t \in [K]$ and then observes $X_t = x_{tA_t}$

Let's look at the worst-case regret

$$R^*_n(\pi) = \sup_{\nu \in [0,1]^{K \times n}} R_n(\pi, \nu)$$

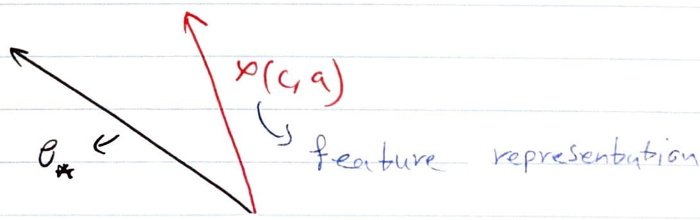
For deterministic policies $\rightarrow$ $R^*_n \geq n(1 - \frac{1}{K})$

Pus 4

There is an algorithm called
Exponential_weight algorithm for
Exploration and Exploitation (EXP3)
which acheives regret of $\leq 2\sqrt{nK \log(K)}$
(chapter 11 _ Bandit Algorithm)

Contextual bandit:
- Consider a bandit setting where at each time step t, we have a set of actions A_t, and, before making a decision, we are also give a context C_t.
When we observe a context c, and choose an action a, we receive a reward X, which depends on a and c

Linear bandit
Consider a setting of Contextual bandit where
$$X_t = \langle \phi(c,a), \theta_* \rangle + \eta_t$$



Def: Let $X_1, \dots$ be a sequence of random variables, on $(\Omega, \mathcal{F}, \mathbb{P})$, and $\mathbb{F} = \{\mathcal{F}_t\}_{t \geq 0}$ a filtration.
An $\mathcal{F}$-adapted sequence of random variable $(X_t)_{t \geq 0}$ is a $\mathcal{F}$-adapted supermartingle sequence if

a) $E[X_t | \mathcal{F}_{t-1}] \leq X_{t-1}$ a.s. $\forall\ t > 0$

b) X_t is integrable $t \geq 0$

Note that a martingale sequence is also supermartingale (similarly, submartingale is defined when $E[X_t | \mathcal{F}_{t-1}] \geq X_{t-1}$)

The general case of maximal inequality:
Theorem: (Maximal inequality): Let $(X_t)_{t \in \mathbb{N}}$ be a supermartingale with $X_t \geq 0$ a.s. for all t.
Then for any $\delta > 0$

$$\mathbb{P}\left(\sup_{t \in \mathbb{N}} X_t > \delta \right) \leq \frac{E[X_0]}{\delta}$$