1) Setting: Linear Contextual Bandits

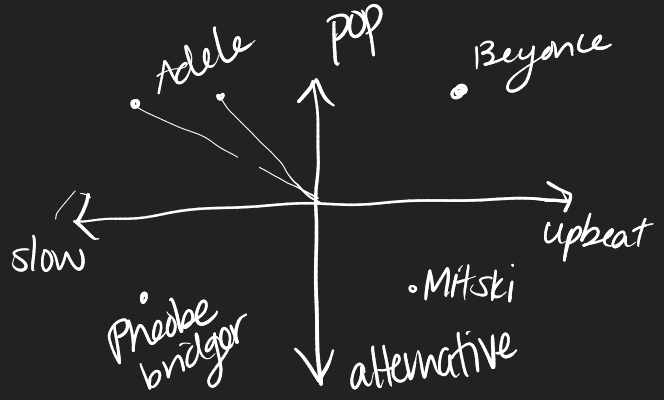
- contexts $x \in \mathcal{X} \subseteq \mathbb{R}^d$ drawn from distribution $\mathcal{D} \in \Delta(\mathcal{X})$ $\quad x_t \overset{iid}{\sim} \mathcal{D}$
- action "arms" $a \in \mathcal{A} = \{1, \dots, K\}$
- rewards $r_t = r(x_t, a_t)$
  $\mathbb{E}[r(x,a)] = \mu_a(x) = \theta_a^T x$ linear function
  $\theta_a \in \mathbb{R}^d$ (unknown)
- horizon T

Goal: find a policy $a_t = \pi(x_t)$, achieve low regret

$$R(T) = \sum_{t=1}^{T} \mathbb{E}_{x_t \sim \mathcal{D}} \Big[\underbrace{\max_a \theta_a^T x_t}_{\mu_{a^*}(x_t)} - \theta_{a_t}^T x_t \Big]$$

$\pi_*(x_t)$

Example: music recommendation
arm = artists
context = user's affinity towards traits



Last lecture: explore-then-commit based on supervised learning

$$a_t = \arg\max_a \hat{\mu}_a(x_t)$$

$$\hat{\mu}_a = \arg\min_{\mu \in \mathcal{M}} \sum_{i=1}^{N} (\mu(x_i^a) - r_i^a)^2$$

data collected during exploration

Linear Regression

since $r_a(x) = \theta_a^T x$, $M = \{ r(x) = \theta^T x \mid \theta \in \mathbb{R}^d \}$

$$\hat{\theta}_a = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} (\theta^T x_i^a - r_i^a)^2$$

Lemma: as long as $(x_i)_{i=1}^N$ span $\mathbb{R}^d$

$$\hat{\theta} = \underbrace{\left(\sum_{i=1}^{N} x_i x_i^T\right)^{-1}}_{A} \underbrace{\sum_{i=1}^{N} x_i r_i}_{b} = A^{-1} b$$

Proof: $\nabla_\theta \left[\sum_{i=1}^{N} (\theta^T x_i - r_i)^2\right] = 2 \sum_{i=1}^{N} (\underbrace{\theta^T x_i}_{x_i^T \theta} - r_i) x_i = 0$

$$\underbrace{\left[\sum_{i=1}^{N} x_i x_i^T\right]}_{A} \theta = \underbrace{\sum_{i=1}^{N} x_i r_i}_{b}$$

□

A is related to empirical covariance of X

$$\Sigma = \mathbb{E}[XX^T] \qquad \hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T \qquad A = N\hat{\Sigma}$$

3) LinUCB Algorithm

We keep track of:

$$A_t^a = \sum_{k=1}^{t} x_k x_k^T \mathbb{1}\{a_k = a\} \qquad b_t^a = \sum_{k=1}^{t} x_k r_k \mathbb{1}\{a_k = a\}$$

$$\hat{\theta}_t^a = (A_t^a)^{-1} b_t^a$$

Alg: LinUCB

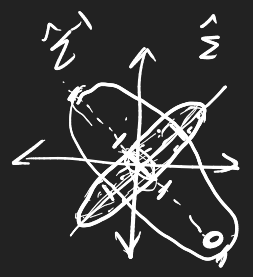
Initialize 0 mean & infinite CI

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

observe x_t

$a_t = \arg\max_a \left[(\hat{\theta}_t^a)^T x_t + \alpha \sqrt{x_t^T (A_t^a)^{-1} x_t} \right]$

observe r_t, update $\hat{\theta}^{a_t}$, b^{a_t}, A^{a_t}



Intuition

$$\underbrace{\hat{\theta}^T x}_{\text{large if } \hat{\theta} \text{ is aligned w/ } x} + \alpha \underbrace{\sqrt{x^T A^{-1} x}}_{\text{large when past data is not aligned with } x}$$

$A = N\hat{\Sigma}$

$x^T A^{-1} x = \frac{1}{N} \underbrace{x^T \hat{\Sigma}^{-1} x}_{\text{large when } x \text{ is } \underline{\text{not}} \text{ aligned with prev. data}}$

$\frac{1}{N}$: amount of data

Statistical Explanation

Claim: With high probability

$$\theta_a^T x \le \hat{\theta}_a^T x + \alpha \sqrt{x^T A_a^{-1} x}$$

where α depends on probability and variance of rewards

Lemma (Chebychev's Inequality)

For a random variable u with $\mathbb{E}[u] = 0$,

$|u| \le \beta \sqrt{\mathbb{E}(u^2)}$ with probability $1 - 1/\beta^2$

Proof of Claim:

We will show, using Chebychev's

$$\underbrace{\left| \hat{\theta}_a^T x - \theta_a^T x \right|}_{= u} \le \alpha \sqrt{x^T A_a^{-1} x}$$ ← from computing $\mathbb{E}[u]$

1) $\underset{r_i}{\mathbb{E}}\left[\hat{\Theta}^T x - \Theta^T x\right] \overset{?}{=} 0$ let $w_i = r_i - \underbrace{\mathbb{E}r_i}_{\Theta^T x_i}$

$$\hat{\Theta} = \left(\sum_{i=1}^N x_i x_i^T\right)^{-1} \sum_{i=1}^N x_i\left(\Theta^T x_i + w_i\right)$$

$$\hat{\Theta} = \left(\sum_{i=1}^N x_i x_i^T\right)^{-1}\left[\left(\sum_{i=1}^N x_i x_i^T\right)\Theta + \sum_{i=1}^N x_i w_i\right]$$

$$= \Theta + \left(\sum_{i=1}^N x_i x_i^T\right)\sum_{i=1}^N x_i w_i$$

$$\mathbb{E}\left[\hat{\Theta}^T x - \Theta^T x\right] = \mathbb{E}\left[(\hat{\Theta} - \Theta)^T x\right] = \mathbb{E}\left[\underbrace{\left(\sum_{i=1}^N x_i x_i^T\right)^{-1}\sum_{i=1}^N x_i w_i}_{r_i - \mathbb{E}r_i}\right]^T x$$

$$\mathbb{E}\left[A^{-1}\sum_{i=1}^N x_i w_i\right] \qquad \mathbb{E}[w_i] = 0$$

$$= A^{-1}\sum_{i=1}^N x_i \mathbb{E}[w_i] \qquad = 0 \quad \checkmark$$