Contextual Bandits & OCB

- ε-Greedy
- SquareCB (inverse ~~pop~~ prop. gap weighting)

Proved given <u>Lemma</u> $\forall \gamma > 0$

$\forall \hat{f}, f^* : [K] \to [0,1]$

$$\pi \sim p(\pi) = \frac{1}{\lambda + \gamma(\hat{f}(\hat{\pi}) - \hat{f}(\pi))}$$

$$\mathbb{E}_{\pi\sim p}\left[f^*(\pi^*) - f^*(\pi)\right]$$

$$\leq \frac{K}{\gamma} + \gamma\,\mathbb{E}\left[\left(\hat{f}(\pi) - f^*(\pi)\right)^2\right]$$

<u>Pf:</u>

$$\mathbb{E}_p\left[f^*(\pi^*) - f^*(\pi)\right]$$

$$= \mathbb{E}_p\Big[f^*(\pi^*) - \hat{f}(\pi^*)$$

$$+ \hat{f}(\pi^*) - \hat{f}(\hat{\pi})$$

$$+ \cdots\Big]$$

$$\leq \frac{K}{\gamma} + \gamma\, \mathbb{E}_p\left[\left(\hat{f}(\pi) - f^*(\pi)\right)^2\right]$$

$\mathbb{E}\,\text{Regret}$ (over T rounds)

$$\leq \frac{KT}{\gamma} + \gamma \underbrace{\sum_{t=1}^{T} \mathbb{E}_p\left[\left(\hat{f}_t(\pi) - f^*(\pi)\right)^2\right]}_{\text{RegretSQ}}$$

$$\gamma = \sqrt{\frac{KT}{\text{RegretSQ}}}$$

$$= 2\sqrt{KT\,\text{RegretSQ}}$$

Today: LinUCB

High-level: (1) Use linear regression ($\vec{\hat{\theta}}_t$) to estimate the reward of each arm

(2) Use uncertainty estimates & $\vec{\hat{\theta}}_t$ to get UCB on each arm

(3) "Optimistic" pick arm w/ best UCB.

Algorithm: $\theta^* \in \mathbb{R}^d$, $\beta > 0$ (Assume $\phi(\cdot,\cdot)$ known

$\|\theta^*\|_2 \le 1$)

$\mathbb{E}[r_t | x_t, \pi_t] = \langle \theta^*, \phi(x_t, \pi_t) \rangle$

For $t = 0$ to $T-1$.

- $$\hat{\theta}_t = \underset{\|\theta\|_2 \le 1}{\operatorname{argmin}} \sum_{s=1}^{t} (r_s - \langle \theta, \underbrace{\phi(x_s, \pi_s)}_{\phi_s} \rangle)^2$$

- Define

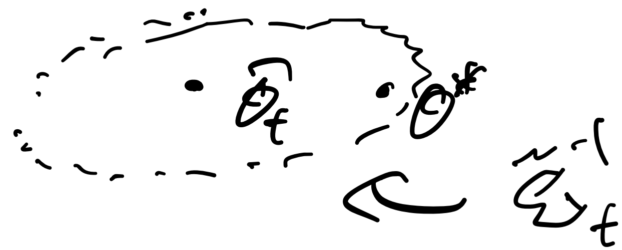
$$\tilde{\Sigma}_t = \sum_{s=1}^{t} \phi_s \phi_s^T + I$$

$$\left(\text{so } \langle u, \tilde{\Sigma}_t u \rangle = \sum_{s=1}^{t} \langle \phi_s, u \rangle^2 + \|u\|_2^2 \right)$$

"how much we know about direction u"

- Observe context x_{t+1}.

Θ_t



- Select

$$\pi_{t+1} \in \underset{\pi \in [K]}{\arg\max} \left[\max_{\theta \text{ s.t.}} \langle \theta, \phi(x_{t+1}, \pi) \rangle \right]$$

"Uncertainty set Θ_t" $=$

$$\langle \theta - \hat{\theta}_t, \tilde{\Sigma}_t (\theta - \hat{\theta}_t) \rangle \le (16\beta + 4)^2$$

- Play π_{t+1}, get reward r_{t+1}.

Possible issues:

- Θ_t too small, doesn't include θ^*
- Θ_t too big.

<u>Thm</u>: $\mathbb{E}\,\text{Regret} = \mathbb{E}\sum_{t=1}^{T}\left(f^*(x_t,\pi_t^*) - f^*(x_t,\pi_t)\right)$

$$\lesssim d\sqrt{T}\log T$$

where $\beta = \Theta(d\log T)$.

<u>Pf</u> Has two parts.

<u>Part I:</u> Show (with high probability) θ^* is always in the uncertainty set.

$$\pi^*_{t+1} = \underset{\pi \in [K]}{\operatorname{argmax}} \langle \theta^*, \phi(x_{t+1}, \pi) \rangle$$

Pf:

$$\mathbb{E}\,\text{Regret} = \sum_{t=0}^{T-1} \left(\langle \theta^*, \phi(x_{t+1}, \pi^*_{t+1}) \rangle - \langle \theta^*, \phi(x_{t+1}, \pi_{t+1}) \rangle \right)$$

Step 1 (<u>w.h.p.</u>) $\theta^* \in \Theta_t$

$$\le \sum_{t=0}^{T-1} \left(\max_{\theta \in \Theta_t} \langle \theta, \phi(x_{t+1}, \pi^*_{t+1}) \rangle - \langle \theta^*, \phi(x_{t+1}, \pi_{t+1}) \rangle \right)$$

$$\le \sum_{t=0}^{T-1} \left(\max_{\theta \in \Theta_t} \langle \theta, \phi(x_{t+1}, \pi_{t+1}) \rangle - \langle \theta^*, \phi(x_{t+1}, \pi_{t+1}) \rangle \right)$$

$$\max \langle \theta - \theta^*, \phi(x_{t+1}, \pi_{t+1}) \rangle$$

$$\max_{\theta} \left\langle \theta - \theta^*, \phi(x_t, \pi_t) \right\rangle$$

$$= \max_{\theta} \left\langle \tilde{\Sigma}_t^{1/2} (\theta - \theta^*), \tilde{\Sigma}_t^{-1/2} \phi(x_t, \pi_t) \right\rangle$$

<u>Lemma</u> (Elliptical Potential Lemma)

Suppose $\phi_1 \dots \phi_T \in \mathbb{R}^d$ $\|\phi_i\| \le 1$

$$\tilde{\Sigma}_t = \sum_{s=1}^{t} \phi_s \phi_s^T + I$$

Then

$$\underbrace{\sum_{t=1}^{T} \phi_t \tilde{\Sigma}_{t-1}^{-1} \phi_t}_{\text{"how surprising } \phi_t \text{ is"}} \leq 2d \log T.$$

Why inverse?

$$Y = X\theta^* + \text{noise}$$

$$X^{-1}Y = \theta^* + X^{-1}\text{noise}$$

(classic OLS)

Completion of pf of Thm, given Lemma.

$$\mathbb{E}\,\mathrm{Regret} \le \mathbb{E} \sum_{t=1}^{T} \max_{\theta \in \Theta_{t-1}} \langle \theta - \theta^*, \phi(x_t, \pi_t) \rangle$$

$$= \mathbb{E} \sum_{t=1}^{T} \max_{\theta \in \Theta_{t-1}} \langle \tilde{\Sigma}_{t-1}^{1/2}(\theta - \theta^*), \tilde{\Sigma}_{t-1}^{-1/2} \phi(x_t, \pi_t) \rangle$$

$$\le \mathbb{E} \sum_{t=1}^{T} \sqrt{\langle \theta - \theta^*, \tilde{\Sigma}_{t-1}(\theta - \theta^*) \rangle} \sqrt{\langle \phi(x_t, \pi_t), \tilde{\Sigma}_t^{-1} \phi(x_t, \pi_t) \rangle}$$

$$\le \sqrt{16\beta^2 + 4} \sqrt{T} \sqrt{\sum_{t=1}^{T} \langle \phi(x_t, \pi_t), \tilde{\Sigma}_{t-1}^{-1} \phi(x_t, \pi_t) \rangle}$$

Pf of Elliptical potential lemma.

Key idea: compute $\det \tilde{\Sigma}_{t+1}$ in terms of $\det \tilde{\Sigma}_t$.

Facts from linear algebra

① $\det(M) = \prod_{i=1}^{d} \lambda_i(M)$

② $\lambda_1(I + uu^T) = 1 + u^T u$

$$\det(\tilde{\Sigma}_{t+1}) = \det\left(\tilde{\Sigma}_t + \phi_{t+1}\phi_{t+1}^T\right)$$

$$= \det\left(\tilde{\Sigma}_t^{1/2}\left(I + \tilde{\Sigma}_t^{-1/2}\phi_{t+1}\phi_{t+1}^T\tilde{\Sigma}_t^{-1/2}\right)\tilde{\Sigma}_t^{1/2}\right)$$

$$= \det(\tilde{\Sigma}_t)\det\left(I + \tilde{\Sigma}_t^{-1/2}\phi_{t+1}\phi_{t+1}^T\tilde{\Sigma}_t^{-1/2}\right)$$

$$= \det(\tilde{\Sigma}_t)\quad\left(1 + \phi_{t+1}^T\tilde{\Sigma}_t^{-1}\phi_{t+1}\right)$$

$$\log\det\tilde{\Sigma}_{t+1} - \log\det\tilde{\Sigma}_t$$

$$= \log\left(1 + \phi_{t+1}^T \tilde{\Sigma}_t^{-1} \phi_{t+1}\right)$$

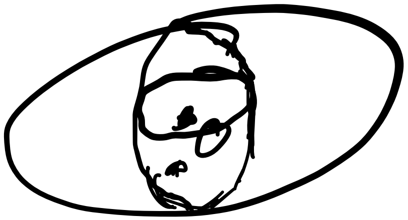
$$\geq \frac{\phi_{t+1}^T \tilde{\Sigma}_t^{-1} \phi_{t+1}}{2}$$

Fact:
$\log(1+x)$ for $x \in [0,1]$
$\geq \frac{x}{2}$

"total amount of surprise"

$$\underbrace{\log\det\tilde{\Sigma}_T}_{\sum_{i=1}^{d} \log\lambda_i(\tilde{\Sigma}_T) \leq d\log(T+1)} - \overbrace{\log\det\tilde{\Sigma}_0}^{\log\det I = 0} \geq \sum_{t=0}^{T-1} \frac{\phi_{t+1}^T \tilde{\Sigma}_t^{-1} \phi_{t+1}}{2}$$

Picture of what happens during Lin UCB

 $\Theta_0 \longleftrightarrow \tilde{\Sigma}_0^{-1}$, $\hat{\Theta}_0 = 0$

@1

$\longrightarrow \phi_1 = e_1 \longrightarrow \tilde{\Sigma}_1 = \tilde{\Sigma}_0 + \phi_1 \phi_1^T$

$\longrightarrow \phi_2 = e_2$

Vol ellipse $\Longleftrightarrow \det \tilde{\Sigma}_t^{-1}$

$= \dfrac{1}{\det \tilde{\Sigma}_t}$

Markov Decision Processes (Intro to RL)

<u>So far</u>: bandit, contextual bandit

Nature

$x_t \longrightarrow$ Decision Maker (pick π)

π_1 $\quad r_1$

π_2 $\quad r_2$

(observe reward)

Regret $\longleftrightarrow$ for the <u>same</u> seq of x_t, how good vs. best decision in hindsight

What is an MDP?

$S = \{\text{state space}\}$

$A = \{\text{action space}\}$

$\Delta(S) = \{\text{probdists over } S\}$

"Transition kernel"

$$P : S \times A \to \Delta(S)$$

"Reward distribution"

$$R : S \times A \to \Delta(\mathbb{R})$$

Initial state s_0 $\xrightarrow{\text{pick action } a_1}$ s_1 $\xrightarrow{a_2}$ Nature samples $\rightarrow$ s_2

$s_2 \sim P(s_1, a_2)$

$r_2 \sim R(s_1, a_2)$

(unpicked action)

Goal: Optimize total reward.