0) Finishing LinUCB Analysis

$$\hat{\theta} = A^{-1}b \qquad A = \sum_{i=1}^{N} x_i x_i^T \qquad b = \sum_{i=1}^{N} x_i \underbrace{r_i}$$

$$r_i = \mathbb{E}[r_i] + w_i$$

last time: $\underset{\text{over noisy rewards}}{\mathbb{E}}\left[\underline{(\hat{\theta} - \theta)^T x}\right] = 0$

$$\hat{\theta} - \theta = \underbrace{A^{-1}\sum_{i=1}^{N} x_i w_i}$$

2) Computing variance

$$\mathbb{E}[((\hat{\theta}-\theta)^T x)^2] = \mathbb{E}\left[x^T\left(\underbrace{A^{-1}\sum_{i=1}^{N} x_i w_i}\right)\left(\underbrace{A^{-1}\sum_{i=1}^{N} x_i w_i}\right)^T x\right]$$

$$= \mathbb{E}\left[x^T A^{-1} \sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j^T w_i w_j A^{-1} x\right]$$

$\mathbb{E}(w_i) = 0$

$$= x^T A^{-1} \sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j^T \,\mathbb{E}[w_i w_j] A^{-1} x$$

let $\sigma^2 = \mathbb{E}(w_i^2)$

$$= x^T \cancel{A^{-1}} \underbrace{\sum_{i=1}^{N} x_i x_i^T}_{\cancel{A}} \sigma^2 A^{-1} x$$

$$= \underline{\sigma^2 x^T A^{-1} x}$$

Chebyshevs:

$$|\underbrace{\hat{\theta}^T x}_{\mu} - \theta^T x| \le \underline{\beta \cdot \sqrt{\sigma^2 x^T A^{-1} x}}$$

UCB: $\theta^T x \le \hat{\theta}^T x + \underbrace{\beta\sigma}_{\alpha}\sqrt{x^T A^{-1} x}$

1) MBRL w/ Exploration

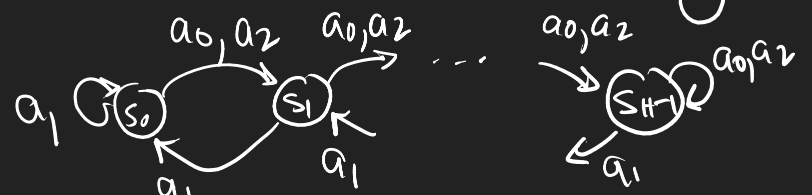
Finite horizon tabular MDP:

$$\mathcal{M} = \{ \mathcal{S}, \mathcal{A}, P, r, H, s_0 \}$$

$|\mathcal{S}| = S$, $|\mathcal{A}| = A$, P unknown

Each episode, we start at s_0 and run forward for H steps. Then reset to s_0 & repeat

example: Need For Strategic Exploration



$$r(s,a) = \begin{cases} 0 & s \neq s_{H-1} \\ 1 & s = s_{H-1} \end{cases}$$

Probability of random walk hitting s_{H-1} starting form s_0 is $(2/3)^H$

Naive idea: MDP $\rightarrow$ MAB

MAB: find the best of K actions

"Tabular" contextual bandits: find the best of K actions for M contexts

$\rightarrow$ KM policies

MDP: find the best policies

2) Upper Confidence Bound Value Iteration

Optimistic Model based RL

Alg: UCB-VI

initiaze guess $\hat{P}_0$ and reward bonus $b_0(s,a)$

For $i = 0, \dots T-1$

optimistically plan: $\pi^i = VI(\hat{P}_i, r+b_i)$

collect new trajectory with $\pi^i = (\pi^i_0, \dots, \pi^i_{H-1})$

update $\hat{P}_{i+1}$ and b_{i+1}

Model Estimation:

$\hat{P}_i$ using dataset $\{\{s^k_t, a^k_t\}_{t=0}^{H-1}\}_{k=0}^{i}$

counts: $$N_i(s,a) = \sum_{k=1}^{i}\sum_{t=0}^{H-1} \mathbb{1}\{(s^k_t, a^k_t) = (s,a)\}$$

$$N_i(s,a,s') = \sum_{k=1}^{i}\sum_{t=0}^{H-1} \mathbb{1}\{(s^k_t, a^k_t, s^k_{t+1}) = (s,a,s')\}$$

$$\hat{P}_i(s'|s,a) = \frac{N_i(s,a,s')}{N_i(s,a)}$$

Reward Bonus: $b_i(s,a) = H\sqrt{\frac{\alpha}{N_i(s,a)}}$

Encourage exploration of new state-action pairs

Generate Policy

In the finite horizon case VI reduces to DP

$\hat{V}^i_H(s) = 0 \quad \forall s$

For $t = H-1, H-2, \dots, 0$

$$\hat{Q}^i_t(s,a) = \underbrace{(r(s,a) + b_i(s,a))}_{r_i(s,a)} + \mathbb{E}_{s' \sim \hat{P}_i(s,a)}[\hat{V}^i_{t+1}(s')]$$

$$\pi^i_t(s) = \arg\max_a \hat{Q}^i_t(s,a)$$

$$\hat{V}^i_t(s) = \hat{Q}^i_t(s, \pi^i_t(s))$$

3) Analysis of UCB-VI

Two key facts:

1) Exploration bonus bounds the difference:

$$\left| \mathbb{E}_{s' \sim \hat{P}_i(s,a)}[V(s')] - \mathbb{E}_{s' \sim P(s,a)}[V(s')] \right| \leq \underline{b_i(s,a)}$$

with high probability.

2) The exploration yeilds <u>optimism</u>

$$\hat{V}_t^i(s) \geq V_t^*(s)$$

Regret Bound:

$= $ best cumulative reward

$\to$ actual cumulative reward

$$R(T) = \mathbb{E}\left[\sum_{i=1}^{T} \underline{V_0^*(s_0) - V_0^{\pi^i}(s_0)} \right]$$

<u>Lemma</u> (Exploration Bonus): For any fixed $V: \mathcal{S} \to [0, H]$ with high probability,

$$\left| \mathbb{E}_{s' \sim \hat{P}_i(s,a)}[V(s')] - \mathbb{E}_{s' \sim P(s,a)}[V(s')] \right| \leq H \sqrt{\frac{\alpha}{N_i(s,a)}} = b_i(s,a)$$

<u>Proof</u>:

$$\left| \mathbb{E}_{s' \sim \hat{P}_i(s,a)}[V(s')] - \mathbb{E}_{s' \sim P(s,a)}[V(s')] \right| = \left| \sum_{s' \in \mathcal{S}} [\hat{P}_i(s'|s,a) - P(s'|s,a)] V(s') \right|$$

$$\leq \sum_{s' \in \mathcal{S}} \left| \hat{P}_i(s'|s,a) - P(s'|s,a) \right| V(s')$$

$$\leq \overbrace{\left(\max_{s'} V(s')\right)}^{\leq H} \sqrt{\frac{\alpha}{N_i(s,a)}}$$

Lemma (optimism): as long as $r(s,a) \in [0,1]$

$$\hat{V}_t^i(s) \geq V_t^*(s) \quad \forall\, t, i, s$$

Proof: Induction $\hat{V}_H^i(s) = 0 = V_H^*(s)$ ✓

Suppose $\hat{V}_{t+1}^i(s) \geq V_{t+1}^*(s)$.

Then: for any s, a

$$\hat{Q}_t^i(s,a) - Q_t^*(s,a) = \cancel{r(s,a)} + b_i(s,a) + \mathbb{E}_{s' \sim \hat{P}_i(s,a)}[\hat{V}_{t+1}^i(s')] - \cancel{r(s,a)} - \mathbb{E}_{s' \sim P(s,a)}[V_{t+1}^*(s')]$$

$$\geq b_i(s,a) - \left| \mathbb{E}_{s' \sim \hat{P}_i(s,a)}[V_{t+1}^*(s')] - \mathbb{E}_{s \sim P(s,a)}[V_{t+1}^*(s')] \right|$$

$$\geq b_i(s,a) - b_i(s,a) = 0$$

$$\hat{Q}_t^i(s,a) \geq Q_t^*(s,a) \quad \forall\, s,a$$

$$\Rightarrow \hat{V}_t^i(s) \geq V_t^*(s) \quad \forall\, s$$