SO FAR: Bandits + Func Approx

MDP + TODAY

QUESTION: Why use function approximation?
State (s) can be very high-dimensional (eg, image)

---

QUESTION: what do we approximate?

RECALL: Linear Bandits

Assumption: $r(s,a) = \phi(s,a)^T \theta^*, \quad \theta^* \in \mathbb{R}^d$

Why? Knowledge of $r \mapsto$ Knowledge of $\pi^*$ (optimal policy)

Notice: Knowledge of $Q^* \mapsto$ Knowledge of $\pi^*$

$$Q_h^*(s,a) = \mathbb{E}_{\pi^*}\left[\sum_{t=h}^{H} r_t \mid (s_h, a_h) = (s,a)\right]$$

---

Is Linear $Q^*$ sufficient to learn sample efficiently?

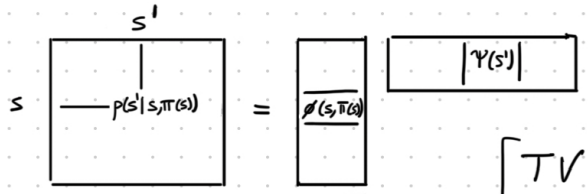
| | (LINEAR) BANDITS | $Q^*$ LINEAR |
|---|---|---|
| UPPER BOUND | $\tilde{O}(d\sqrt{K})$ (LINUCB) ← # episodes | $O(e^H)$ |
| LOWER BOUND | $\Omega(d\sqrt{K})$ | $\Omega(e^d)$ |

NO ✗ => Need additional assumptions!

## LINEAR MDPS $\approx$ MDPS that can be decomposed as a sequence of Linear Bandits

$$\text{Def}: \begin{cases} r_h(s,a) = \phi(s,a)^\top \theta_h^R \\ P_h(s'|s,a) = \phi(s,a)^\top \psi(s') \end{cases}$$

Also called "Low Rank" MDPs:




$$\forall \pi: \quad P^\pi = \Phi^\pi \psi^\top \qquad \left[ TV = R + PV = (\Phi\theta^R + \Phi\psi V) \in \text{Range } \Phi \right]$$

## FOUNDAMENTAL PROPERTY OF LINEAR MDPS

**Def:** Bellman operator $(T_h V)(s,a) = r_h(s,a) + \sum_{s'} p_h(s'|s,a) V(s')$

**PROPOSITION** (← Bellman operator)

For any value function $V$, $(T_h V)(s,a) = \phi(s,a)^\top \theta$ for some $\theta$.

Proof:

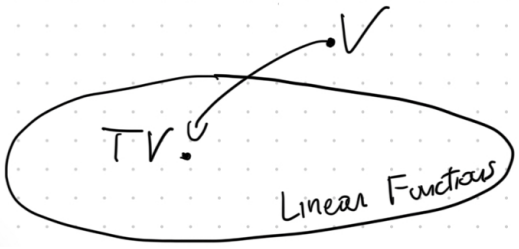
$$(T_h V)(s,a) \overset{\text{def}}{=} r_h(s,a) + \sum_{s'} p(s'|s,a) V(s')$$

$$= \phi(s,a)^\top \theta_h^R + \sum_{s'} \phi(s,a) \psi(s') V(s')$$

$$= \phi(s,a)\Big[\theta_h^R + \underbrace{\sum_{s'} \psi(s') V(s')}_{\text{def}}\Big]$$

$$= \mathring{\theta}(V) \in \mathbb{R}^{d'}$$

$$= \phi(s,a)^\top (\theta_h^R + \mathring{\theta}(V))$$



# LEARNING EFFICIENTLY ON LINEAR MDPS

## ALGORITHM: LEAST-SQUARE VALUE ITERATION WITH UCB

for episode $k = 1, \dots, K$ $\qquad D = \{ (s_{tk}, a_{tk}, r_{tk}, s'_{t+1,k}) \}_{t=1,\dots,H}^{k=1,\dots,K}$ ← Dataset

for timestep $h = H, \dots, 1$

covariance → $\Sigma_t = \sum_{i=1}^{K-1} \phi_{ti} \phi_{ti}^T + \lambda I$ (covariance)

$\hat{\theta}_t = \Sigma_t^{-1} \sum_{i=1}^{K-1} \phi_{ti} [ r(s_{ti}, a_{ti}) + \max_{a'} \bar{Q}_{t+1}(s_{t+1,i}, a') ]$

$\bar{Q}_t(\cdot,\cdot) = \min \{ H, \phi(\cdot,\cdot)^T \hat{\theta}_t + \underbrace{\sqrt{\beta} \| \phi(\cdot,\cdot) \|_{\Sigma_t^{-1}}}_{\text{bonus } b_t(\cdot,\cdot)} \}$

end for

Deploy $\pi$ and add trajectory to the dataset

---

## FIRST STEP ANALYSIS ≈

Total Error at timestep $(t)$ = estimation error + bonus + next-state error

$$(\bar{Q}_{tk} - Q_t^\pi)(s,a) = \phi(s,a)^T \hat{\theta}_{tk} + b_t(s,a) - r(s,a) - E_{s' \sim p(s,a)} V_{t+1}^\pi(s')$$

$$\hat{\theta}_{tk} = \Sigma_{tk}^{-1} \left[ \sum_{i=1}^{K-1} \phi_{ti} \left\{ r_{ti} + \max_{a'} \bar{Q}_{t+1}(s_{t+1,i}, a') \right\} \right]$$

$$= E_{s' \sim p(s_{ti}, a_{ti})} \max_{a'} \bar{Q}_{t+1}(s', a') + \underbrace{\eta_{ti}}_{\text{(transition noise)}}$$

$$\left\{ \phi_{ti}^T \theta_t^R + \phi_{ti}^T \theta(\bar{V}_{t+1,i}) + \eta \right\} \quad \text{(Linear MDP Property)}$$

$$= \underbrace{\phi(s,a)^T \Sigma_{tk}^{-1} \sum_{tk} \left[ \theta_t^R + \theta(\bar{V}_{t+1,i}) \right]}_{\text{def } e_t(s,a) \text{ (estimation error)}} + \boxed{\phi(s,a) \Sigma_{tk}^{-1} \sum_{i=1}^{K-1} \phi_{ti} \eta_{ti}} - r(s,a) - E_{s' \sim p(s,a)} V^\pi(s') + b_t(s,a)$$

$$= \cancel{r(s,a)} + E_{s' \sim p(s,a)} \bar{V}_{t+1,i}(s') - \cancel{r(s,a)} - E_{s' \sim p(s,a)} V_{t+1}^\pi + e_t(s,a) + b_t(s,a)$$

$$= E_{s' \sim p(s,a)} [ \bar{V}_{t+1,i} - V_{t+1}^\pi ](s') + b_t(s,a) + e_t(s,a)$$

## BOUND ON TRANSITION ERROR AND OPTIMISM

LEMMA

$\sqrt{\beta} = \tilde{\mathcal{O}}(dH)$

w.p. $\geq 1-\delta$: $|e_t(s,a)| = \phi(s,a)^\top \Sigma_{tk}^{-1} \sum_{i=1}^{K-1} \phi_{ti} y_{ti} \leq \|\phi(s,a)\|_{\Sigma_{tk}^{-1}} \|\sum_{i=1}^{K-1} \phi_{ti} y_{ti}\|_{\Sigma_{tk}^{-1}}$

$$\left(\bar{Q}_{tk} - Q^{\pi^*}\right)(s,\pi^*(s)) = \underbrace{E_{s' \sim p(s,\pi^*(s))}\left[\bar{V}_{t+1,k} - V_{t+1}^{\pi^*}\right](s')}_{\geq 0 \text{ by induction}} + \underbrace{b_t(s,a) - e_t(s,a)}_{\text{(this is how the bonus is chosen)}} \geq 0$$

---

## REGRET BOUND

$$\sum_{k=1}^{K}\left[V^* - V^{\pi_k}\right] \leq \sum_{k=1}^{K}\left[\bar{V}_k - V^{\pi_k}\right] = \sum_{k}^{K}\sum_{t=1}^{H}\Big[\underbrace{b_{tk}}_{\sqrt{\beta}\|\phi_{tk}\|_{\Sigma_{tk}^{-1}}}(s_{tk}, a_{tk}) + \underbrace{e_{tk}(s_{tk}, a_{tk})}_{\text{same order}}\Big]$$

(experienced state-actions)

$$\overset{CS}{\leq} \sum_{t=1}^{H} \sqrt{K}\sqrt{\beta}\sqrt{\underbrace{\sum_k \|\phi_{tk}\|^2_{\Sigma_{tk}^{-1}}}_{\mathcal{O}(d)}}$$

$$\sim H\sqrt{K}\underbrace{\sqrt{\beta}}_{\tilde{\mathcal{O}}(dH)}\sqrt{d} \sim H^{\frac{3}{2}} d^{\frac{1}{2}} \sqrt{T}$$

($T$ = # steps)