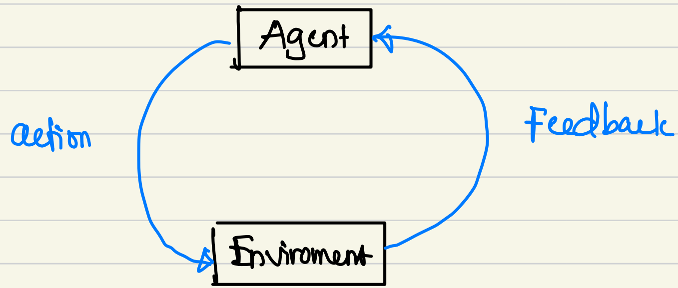
June 28, 23

Unit 1: MDP basics

MDP formulations



Markov Decision process (MDP)

Temporal correlation: actions has influence on the futures

focus on this for simplicity

- Infinite-horizon (discounted) MDP: $M = (S, A, P, r, \gamma)$
- Finite-horizon MDP: $M = (S, A, \{P_h\}_{h \in [H]}, \{r_h\}_{h \in [H]}, H)$

Interaction protocol: (Discounted MDP)

- Start: $s_0 \sim \mu$
- for time step $t = 1, 2, \ldots, T$:
  - agent
    - takes $a_t \in \mathcal{A}$
    - obtains rewards: $r_t = r(s_t, a_t)$
    - observes next state: $s_{t+1} \sim P(\cdot \mid s_t, a_t)$

$\mathcal{S}$: State space

$\mathcal{A}$: action space

$r: \mathcal{S} \times \mathcal{A} \longrightarrow [0,1]$ reward function

$P: \mathcal{S} \times \mathcal{A} \longrightarrow \Delta(\mathcal{S})$, transition probability

$P(s' \mid s, a)$

γ: discount factor

Environment:

P, r

① known — Dynamic Programming

② simulator — "generative model"

③ unknown, have to play from beginning — RL

Policy:

History-dependent policies: $\Pi^{hist} = \{ (S \times A \times \mathbb{R})^* \times S \longrightarrow \Delta(A) \}$

Stationary policies: $\Pi^{stn} = \{ S \longrightarrow \Delta(A) \}$

Deterministic policies: $\Pi^{det} = \{ S \longrightarrow A \}$

Value: $V_M^\pi(s)$: cumulative rewards

$$V_M^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \pi, s_0 = s\right]$$

$$Q_M^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \pi, s_0 = s, a_0 = a\right]$$

Goal: Find an optimal policy π^*, i.e.,

$$V_M^{\pi^*}(s) \geq V_M^\pi(s) \quad \forall (s, \pi) \in (\mathcal{S}, \Pi^{\text{hist}})$$

Theorem: $\boxed{\exists\, \pi^* \in \Pi^{det}}$

Bellman equations: $\forall \pi \in \Pi^{stn}$

$$\begin{cases} V_M^{\pi}(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} \left[Q_M^{\pi}(s,a) \right] \\ Q_M^{\pi}(s,a) = r(s,a) + \gamma \, \mathbb{E}_{\substack{a \sim \pi(\cdot|s) \\ s' \sim P(\cdot|s,a)}} \left[V^{\pi}(s') \right] \end{cases}$$

Bellman optimality equations:

$$\begin{cases} V_M^{\pi^*}(s) = \max_{a \in A} Q_M^{\pi^*}(s,a) \\ Q_M^{\pi^*}(s,a) = r(s,a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a'} Q_M^{\pi^*}(s',a') \right] \end{cases}$$

Bellman optimality operator: $T: \{S \times A \to \mathbb{R}\} \longrightarrow \{S \times A \to \mathbb{R}\}$

$$[TQ](s,a) = r(s,a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a'} Q(s',a') \right]$$

Unit 2: MDP Planning

known r, P

(consider $\pi \in \Pi^{stn}$)

- Value iteration
- Policy iteration
- Linear programming

Notations: • If context is clear:

$$V_M^\pi \longrightarrow V^\pi, \quad Q_M^\pi \longrightarrow Q^\pi, \quad V_M^{\pi^*} \longrightarrow V^*, \quad Q_M^{\pi^*} \longrightarrow Q^*$$

• Vectorization everything.

$$V^\pi = [V^\pi(s)]_{s \in S} \in \mathbb{R}^{|S|}, \quad Q^\pi \in \mathbb{R}^{|S|.|A|}$$

$$r \in \mathbb{R}^{|S|.|A|}, \quad P = [P(s'|s,a)]_{s',s,a} \in \mathbb{R}^{|S|.|A| \times |S|}$$

$$P^\pi = \left[P^\pi_{(s',a'),(s,a)} \right] \in \mathbb{R}^{|S||A| \times |S|.|A|}$$

$$P^\pi_{(s',a'),(s,a)} = P(s'|s,a)\, \pi(a'|s').$$

Greedy policy: Given $Q \in \{S \times A \to \mathbb{R}\}$

$$\pi_Q(s) \in \underset{a \in A}{\operatorname{argmax}}\; Q(s,a), \; \forall s$$

vectorized Bellman equations

- $$Q^\pi = r + \gamma P V^\pi$$
$$= r + \gamma P^\pi Q^\pi$$

claim : $$\boxed{Q^\pi = (I - \gamma P^\pi)^{-1} r}$$

proof : - only need to prove $I - \gamma P^\pi$ is invertible

- $\forall x \in \mathbb{R}^{|S|}$, s.t. $\|x\|_\infty > 0$,

$$\|(I - \gamma P^\pi)x\|_\infty = \|x - \gamma P^\pi x\|_\infty$$
$$\geq \|x\|_\infty - \gamma \|P^\pi x\|_\infty$$
$$\geq \|x\|_\infty - \gamma \|x\|_\infty$$
$$= (1-\gamma)\|x\|_\infty > 0$$

$\because$ row $(P^\pi x)$ is an average of the coordinates of x.

Claim: $(1-\gamma)(I-\gamma P^{\pi})^{-1} = \underbrace{(1-\gamma)\sum_{t=0}^{\infty} \gamma^t (P^{\pi})^t}_{\text{a mixture of distributions}}$

Claim (contraction):

$$\|TQ - TQ'\|_\infty \leq \gamma \|Q - Q'\|_\infty$$

$$\forall Q, Q' \in \{S \times A \to \mathbb{R}\}$$

Value iteration:

- set $Q_0 = 0^{|S|\cdot|A|}$
- iterate: $Q_{t+1} = TQ_t \,, \quad t = 0, 1, \ldots$

Claim

$$\|Q_t - Q^*\|_\infty \leq \gamma^t \|Q^*\|_\infty$$

Policy iteration:

- start from an arbitrary policy π_0
- For $t = 0, 1, \ldots, :$
  - Policy evaluation: Q^{π_t}
  - Greedify: $\pi_{t+1} = \pi_{Q^{\pi_t}}$

Claim:

- $Q^{\pi_t} \leq TQ^{\pi_t} \leq Q^{\pi_{t+1}}$
- $\|Q^{\pi_{t+1}} - Q^*\|_\infty \leq \gamma \|Q^{\pi_t} - Q^*\|_\infty$

Linear programming:

Primal:

$$\min_{V \in \mathbb{R}^{|S|}} V^T \mu$$

subject to

$$\left\{ V(s) \geq r(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s') \quad \forall s \in S, a \in A \right.$$

Claim: V^* is the unique solution

Dual LP:

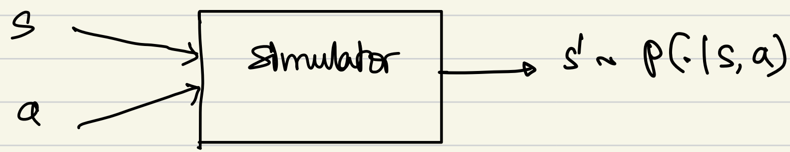
$$\arg\max \frac{1}{1-\gamma} d^T r$$

subject to: $$\sum_{a \in A} d(s,a) = (1-\gamma)\mu(s) + \gamma \sum_{s',a'} P(s|s',a') d(s',a')$$

$$\forall s \in S$$

Unit 3: Simulator setting
(Generative models)

Simulator:



assume r is known to learner

Goal:

- Find ε-optimal value function $\hat{Q}$: (value guarantee / bound)

$$\| Q^* - \hat{Q} \|_\infty \le \varepsilon$$

- Find ε-optimal policy $\hat{\pi}$: (policy guarantee / bound)

$$\| Q^* - Q^{\hat{\pi}} \|_\infty \le \varepsilon$$

more subtle

Model-based method

- maximum likelihood estimate (MLE) of the transition kernel and use it as a "plug-in"

- For each $(s,a) \in S \times A$, draw N samples of $s' \sim P(\cdot|s,a)$

- Let $\hat{P}$ be the empirical model, i.e.

$$\hat{P}(s'|s,a) = \frac{\text{count}(s', s, a)}{N}$$

- #samples $= |S| \cdot |A| \cdot N$

Notations

- Denote $\hat{P} = \left[\hat{P}(s'|s,a) \right]_{(s,a,s')} \in \mathbb{R}^{|S|\cdot|A| \times |S|}$

- Let $\hat{M} = (S, A, r, \hat{P}, \gamma, \mu)$
- Let $\hat{Q}^{\pi} = Q^{\pi}_{\hat{M}}$, $\hat{\pi}$ is an optimal policy of $\hat{M}$
- Let $H = \dfrac{1}{1-\gamma}$: (effective) horizon

Coarse analysis

linear in model complexity

Theorem $\forall \varepsilon, \delta \geq 0$ if $\#\text{samples} = |S|\cdot|A|\cdot N \gtrsim \dfrac{\gamma^2 H^4 |S|^2 \cdot |A| \log\left(\frac{|S|\cdot|A|}{\delta}\right)}{\varepsilon^2}$

- w.p.a.l. $1-\delta$, $\|Q^* - \hat{Q}^*\|_\infty \leq \varepsilon$

- wpal $1-\delta$, $\|Q^* - Q^{\hat{\pi}}\|_\infty \lesssim \varepsilon$

Proof:

$$\|Q^* - \hat{Q}^*\|_\infty = \|\max_\pi Q^\pi - \max_\pi \hat{Q}^\pi\|_\infty$$

$$\leq \underbrace{\max_\pi}_{\text{uniform convergence}} \|\underbrace{Q^\pi - \hat{Q}^\pi}_{}\|_\infty$$

simulation lemma + concentration

uniform convergence

$$\|Q^* - Q^{\hat{\pi}}\|_\infty = \|Q^* - \hat{Q}^* + \hat{Q}^* - Q^{\hat{\pi}}\|_\infty$$
$$\leq \|Q^* - \hat{Q}^*\|_\infty + \|\hat{Q}^{\hat{\pi}} - Q^{\hat{\pi}}\|_\infty$$
$$\leq 2 \cdot \max_{\pi} \|Q^\pi - \hat{Q}^\pi\|_\infty$$

<u>Simulation lemma:</u>

$$Q^\pi - \hat{Q}^\pi = \gamma (I - \gamma \hat{P}^\pi)^{-1} (P - P) V^\pi$$

<u>proof:</u> Recall:
$$Q^\pi = r + \gamma P^\pi Q^\pi \Rightarrow \begin{cases} Q^\pi = (I - \gamma P^\pi)^{-1} r \\ r = (I - \gamma P^\pi) Q^\pi \end{cases}$$

$$Q^\pi - \hat{Q}^\pi = Q^\pi - (I - \gamma \hat{P}^\pi)^{-1} r$$
$$= (I - \gamma \hat{P}^\pi)^{-1} \left((I - \gamma \hat{P}^\pi) Q^\pi - (I - \gamma P^\pi) Q^\pi \right)$$
$$= \gamma (I - \gamma \hat{P}^\pi)^{-1} (P^\pi - \hat{P}^\pi) Q^\pi$$
$$= \gamma (I - \gamma \hat{P}^\pi)^{-1} (P - \hat{P}) V^\pi$$

$$\|Q^\pi - \hat{Q}^\pi\|_\infty = \|\gamma (I - \gamma \hat{P}^\pi)^{-1} (P - \hat{P}) V^\pi\|_\infty$$

$$\leq \frac{\gamma}{1-\gamma} \|(P-\hat{P}) V^\pi\|_\infty$$

$$= \frac{\gamma}{1-\gamma} \max_{(s,a)} \left| \left(P(\cdot|s,a) - \hat{P}(\cdot|s,a)\right)^T V^\pi(\cdot) \right|$$

Hölder

$$\leq \frac{\gamma}{1-\gamma} \max_{(s,a)} \|P(\cdot|s,a) - \hat{P}(\cdot|s,a)\|_1 \cdot \|V^\pi\|_\infty$$

$$\leq \frac{\gamma}{(1-\gamma)^2} \max_{(s,a)} \|P(\cdot|s,a) - \hat{P}(\cdot|s,a)\|_1$$

Using McDiarmid's inequality :

$$\|P(\cdot|s,a) - \hat{P}(\cdot|s,a)\|_2 \leq \sqrt{\frac{2\log(e|S|\cdot|A|/\delta)}{N}}$$

Holder's inequality:

$$\| P(\cdot|s,a) - \hat{P}(\cdot|s,a) \|_1 \leq \sqrt{|S|} \cdot \| P(\cdot|s,a) - \hat{P}(\cdot|s,a) \|_2$$

Crude value bounds

- \# samples in the coarse analysis is linear in model complexity
- it comes from guarantees uniformly over all policies
- yet we only need to care about π^* and $\hat{\pi}$.

- Can # samples be sublinear in model complexity ?