CS292F StatRL Lecture 2
Markov Decision Processes

Instructor: Yu-Xiang Wang

Spring 2021

UC Santa Barbara

Recap: Markov Decision processes (MDP) parameterization

rolled out

$H = (S_1, A_1, R_1, S_2, A_2, R_2, \ldots$

$\sim \mu_0$ $\quad A_1 \sim \pi(a|S=S_1)$

$S_2 \sim P(s|S_1=s, A_1=A_1)$

- Infinite horizon / discounted setting

$$\mathcal{M}(\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu)$$

Transition kernel: $P: S \times A \to \Delta(S)$ i.e. $P(s'|s,a)$

(Expected) reward function: $r: S \times A \to \mathbb{R} / [0, R_{max}]$ $\quad [0,1]$ $\quad \mathbb{E}[R_t \mid S_t=s, A_t=a] =: r(s,a)$

Initial state distribution $\mu_0 \in \Delta(S)$

Discounting factor: $0 \le \gamma \le 1$

Horizon $\frac{1}{1-\gamma} = 1 + \gamma + \gamma^2 + \ldots$

Recap: Reward function and Value functions

- Immediate reward function r(s,a,s')
  - expected **immediate** reward
    $$r(s, a, s') = \mathbb{E}[R_1 | S_1 = s, A_1 = a, S_2 = s']$$
    $$r^\pi(s) = \mathbb{E}_{a \sim \pi(a|s)}[R_1 | S_1 = s]$$
- state value function: $V^\pi(s)$
  - expected **long-term** return when starting in *s* and following π
    $$V^\pi(s) = \mathbb{E}_\pi[R_1 + \gamma R_2 + \ldots + \gamma^{t-1} R_t + \ldots | S_1 = s]$$
- state-action value function: $Q^\pi(s,a)$
  - expected **long-term** return when starting in *s*, performing *a*, and following π

$$Q^\pi(s, a) = \mathbb{E}_\pi[R_1 + \gamma R_2 + \ldots + \gamma^{t-1} R_t + \ldots | S_1 = s, A_1 = a]$$

Recap: Optimal value function and the MDP planning problem

$$V^\star(s) := \sup_{\pi \in \Pi} V^\pi(s)$$

$$Q^\star(s,a) := \sup_{\pi \in \Pi} Q^\pi(s,a).$$

Goal of MDP planning:

$$\text{Find } \pi^* \text{ such that } V^{\pi^*}(s) = V^*(s) \quad \forall s$$

Approximate solution:

$$\pi \text{ is } \epsilon\text{-optimal if } V^\pi \geq V^*(s) - \epsilon \mathbf{1}$$

Recap: General policy, Stationary policy, Deterministic policy

- General policy could depend on the entire history

$$\pi : (\mathcal{S} \times \mathcal{A} \times \mathbb{R})^* \times \mathcal{S} \to \Delta(\mathcal{A})$$

memoryless

- Stationary policy

$$\pi : \mathcal{S} \to \Delta(\mathcal{A})$$

- Stationary, Deterministic policy

$$\pi : \mathcal{S} \to \mathcal{A}$$

Recap: We showed the following results about MDPs.

- **Proposition:** It suffices to consider stationary policies.
  1. Occupancy measure

$$V_\mu^\pi(s) = \sum_{t=1}^{\infty} \gamma^{t-1} \cdot d^\pi(S_t = s)$$

$$V_\mu^\pi(s,a) = \sum_{t=1}^{\infty} \gamma^{t-1} \cdot d^\pi(S_t = s, A_t = a)$$

$$V^\pi(\mu) = \langle V_\mu^\pi(s,a), r(s,a) \rangle$$

$\exists\ \pi'$ stationary s.t. $V^\pi(s,a) = V^{\pi'}(s,a)$

  2. There exists a stationary policy with the same occupancy measure

- **Corollary:** There is a stationary policy that is optimal for all initial states.
  - Proof sketch: 1. Construct an optimal non-stationary policy. 2. Apply the above proposition.

Bellman equations – the fundamental equations of MDP and RL

- For stationary policies there is an alternative, recursive and more useful way of defining the V-function and Q function

$$V^\pi(s) = \overbrace{\sum_a \pi(a|s)}^{\mathbb{E}_{a\sim\pi(a|s)}} \underbrace{\overbrace{\sum_{s'}}^{\mathbb{E}_{s'\sim P(\cdot|s,a)}} P(s'|s,a)[\underbrace{r(s,a,s')}_{\text{immediate reward}} + \underbrace{\gamma}_{\text{discounted}} \underbrace{V^\pi(s')}_{\text{future reward under } \pi}]}_{Q^\pi(s,a)} = \sum_a \pi(a|s) Q^\pi(s,a)$$

- Exercise:
  - Prove Bellman equation from the (first principle) definition.
  - Write down the Bellman equation using Q function alone.

$$Q^\pi(s,a) = ? \; \sum_{s'} P(s'|s,a)\Big[r(s,a,s') + \gamma \underbrace{\sum_{a'} \pi(a'|s')\, Q^\pi(s',a')}_{V^\pi(s')}\Big]$$

Deriving Bellman Equation for stationary policies

Law of total expectation

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$$

$$V^{\pi}(s) = \mathbb{E}^{\pi}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r(S_t, A_t) \,\middle|\, S_1 = s\right]$$

Law of Total expectation

$$= \mathbb{E}^{\pi}[r(S_1, A_1) | S_1 = s] + \underbrace{\mathbb{E}^{\pi}}_{\sum_{s_2} P^{\pi}(S_2=s'|S_1=s)}\left[\sum_{t=2}^{\infty} \gamma^{t-1} r(S_t, A_t) \,\middle|\, S_2 = s'\right] \qquad \tilde{t} = t-1$$

$$= r^{\pi}(s) + \gamma \sum_{s_2} P^{\pi}(S_2=s'|S_1=s)\, \mathbb{E}^{\pi}\left[\sum_{\tilde{t}=1}^{\infty} \gamma^{\tilde{t}-1} r(S_{\tilde{t}}, A_{\tilde{t}}) \,\middle|\, S_1 = s'\right]$$

By stationarity

$$= r^{\pi}(s) + \gamma \sum_{s_2} P^{\pi}(S_2=s'|S_1=s)\, V^{\pi}(s')$$

$$\boxed{V^{\pi} = r^{\pi} + \gamma P^{\pi} \cdot V^{\pi}}$$

$V^{\pi} \in R^{S}$, $P^{\pi} \in R^{S \times S}$

$$\underbrace{P^{\pi}(s'|s)}_{R^{S\times S}} = \sum_{a} \underbrace{P(s'|s,a)}_{R^{S \times SA}} \cdot \pi(a|s)$$

$$P^{\pi} = (\text{Transition Matrix}^{\pi})^{T}$$

Bellman equations in matrix forms

- Lemma 1.4 (Bellman consistency): For stationary policies, we have

$$V^\pi(s) = Q^\pi(s,\pi(s)). = \mathbb{E}_{a\sim\pi(a|s)}\left[Q^\pi(s,a)\right]$$

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

- In matrix forms:

marginalized $\mathbb{R}^{S\times S}$

$$V^\pi = r^\pi + \gamma P^\pi V^\pi \iff (I-\gamma P^\pi)V^\pi = r^\pi \in \mathbb{R}^S$$

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

joint over a'

$$Q^\pi = r + \gamma P^\pi Q^\pi. \qquad (I-\gamma P^\pi)Q^\pi = r \in \mathbb{R}^{SA}$$

$\in \mathbb{R}^{SA \text{ by } SA}$

Closed-form solution for solving for value functions

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

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

$$Q^\pi = r + \gamma P^\pi Q^\pi \,.$$

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

$V^\pi(\mu) = \sum_{s,a} r(s,a)\, \nu_\mu^\pi(s,a)$

$= \langle r, \underline{\nu_\mu^\pi} \rangle$

Duality between *value functions* and *occupancy measures*

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

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

$$Q^\pi = r + \gamma \overset{\mu}{P^\pi} Q^\pi \, .$$

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

$\nu^\pi = (I - \gamma (P^\pi)^\top)^{-1} \mu$

$\nu^\pi(s) = \mu(s) + \gamma \sum_{s'} \nu^\pi(s') \cdot P^\pi(s|s')$

$\nu^\pi = \mu + \gamma (P^\pi)^\top \nu^\pi$

$\tilde{\nu}^\pi = \mu^\pi + \gamma (\tilde{P}^\pi)^\top \tilde{\nu}^\pi$

$\nu^\pi(s,a) = \underbrace{\mu(s) \cdot \pi(a|s)}_{\mu^\pi(s,a)} + \gamma \sum_{s'} \nu^\pi(s') \left[\sum_{a'} P^\pi(s|s',a') \pi(a'|s') \right]$ $\pi(a|s)$

$\nu^\pi(s,a) = \mu^\pi(s,a) + \gamma \sum_{s'} \sum_{a'} \nu^\pi(s',a') \underbrace{P^\pi(s,a|s',a')}_{R^{SA \times SA}}$

A is full rank $\Leftrightarrow$ A^T is full rank

Invertibility of the matrix $I - \gamma P^\pi$

Corollary 1.5 in AJKS: the matrix $I - \gamma P^\pi$ is full rank / invertible for all gamma < 1.

Proof:

identity

$$\begin{aligned}
\|(I - \gamma P^\pi)x\|_\infty &= \|x - \gamma P^\pi x\|_\infty \\
&\geq \|x\|_\infty - \gamma\|P^\pi x\|_\infty \\
&\geq \|x\|_\infty - \gamma\|x\|_\infty \\
&= (1-\gamma)\|x\|_\infty > 0
\end{aligned}$$

triangular inequality

$\gamma < 1$

$$P^\pi x = \begin{pmatrix} \langle P^\pi[1,:], x\rangle \\ \langle P^\pi[2,:], x\rangle \\ \vdots \end{pmatrix}$$

$$\begin{aligned}
&\langle P^\pi[1,:], x\rangle \\
&\leq \|P^\pi[1,:]\|_1 \|x\|_\infty \\
&\leq \|x\|_\infty
\end{aligned}$$

Bellman optimality equations characterizes the optimal policy

expected immediate reward

$$V^*(s) = \max_{a} \sum_{s'} P(s'|s,a)[r(s,a,s') + \gamma V^*(s')]$$

π

discounted future reward by the optimal policy

- system of n non-linear equations
- solve for V*(s)
- easy to extract the optimal policy

- having Q*(s,a) makes it even simpler

$$\pi^*(s) = \arg\max_a Q^*(s,a)$$

Proposition: There is a *deterministic*, *stationary* and *optimal* policy.

- And it is given by:

$$\pi^*(s) = \arg\max_a Q^*(s,a)$$

- Proof:

π^* is stationary

$V^{\tilde{\pi}}(s) \leq V^*(s) = V^{\pi^*}(s) = E_{a \sim \pi^*(a|s)} Q^{\pi^*}(s,a) \leq \max_a Q^{\pi^*}(s,a)$

$\forall \tilde{\pi}$

$= \max_a Q^*(s,a) = Q^*(s, \pi'(s))$

define $\pi'(s) = \arg\max_a Q^*(s,a)$

π' is stationary

π' is deterministic

$= V^{\pi'}(s)$

Substitute $\tilde{\pi} = \pi'$

The crux of solving the MDP planning problem is to construct Q*

- In the remainder of this lecture, we will talk about two approaches

1. By solving a Linear Program

2. By solving Bellman equations / Bellman optimality equations.

The linear programming approach

$V^{\pi^*}(\mu)$ (Ye, 1990s)

- Solve for V* by solving the following LP

$$\min_{V \in \mathbb{R}^S} \quad \sum_s \mu(s) V(s)$$

Substitue $V = V^*$ $\quad \sum_s \mu(s) V^*(s) = V^*(\mu)$

$$\text{subject to} \quad V(s) \geq r(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s') \quad \forall a \in \mathcal{A},\ s \in \mathcal{S}$$

$$V(s) \geq \max_a r(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s')$$

The linear programming approach

- Solve for V* by solving the following LP

$$\min \quad \sum_s \mu(s) V(s)$$

$$\text{subject to} \quad V(s) \geq r(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s') \quad \forall a \in \mathcal{A},\, s \in \mathcal{S}$$

Quiz 1: Once we have V*, how to construct Q*?

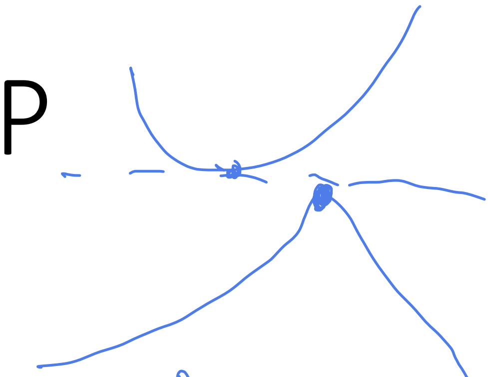
$\pi^*(s) = \arg\max_a Q^*(s,a)$

$Q^*(s,a) = [r(s,a) + \gamma \sum P(s'|s,a) \cdot V^*(s')]$

The Lagrange dual of the LP

$$\max_{\nu} \quad \sum_{s,a} \nu(s,a) r(s,a) \quad = V^*(\mu)$$

$\nu = \nu^{\pi^*}$



$\min_{V} \text{Primal} = \max_{\nu} \text{Dual}$

$$\text{subject to} \quad \nu \geq 0$$

$$\sum_{z} \nu(s,a) = \mu(s) + \gamma \sum_{s',a'} P(s|s',a')\nu(s',a')$$

∇G

- Exercise: Deriving the dual by applying the standard procedure.

The Lagrange dual of the LP

$$\max_{\nu} \quad \sum_{s,a} \nu(s,a) r(s,a)$$

$$\text{subject to} \quad \nu \geq 0$$

$$\sum_{z} \nu(s,a) = \mu(s) + \gamma \sum_{s',a'} P(s|s',a')\nu(s',a')$$

$\nu \in \mathbb{R}^{SA}$

- Exercise: Deriving the dual by applying the standard procedure.

$$\pi^*(a|s) = \frac{\nu^{\pi^*}(s,a)}{\sum_a \nu^{\pi^*}(s,a)}$$

Quiz 2: Once we have the solution how to construct the policy?

$$\nu^*(s,a) = \nu^{\pi^*}(s,a) = \nu^{\pi^*}(s)\cdot \pi^*(a|s)$$

Value iterations for MDP planning

- Recall: Bellman optimality equations

$$V^*(s) = \max_a \sum_{s'} P(s'|s,a)[r(s,a,s') + \gamma V^*(s')]$$

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

$$Q' \leftarrow \mathcal{T}Q = r + PV_Q \quad \text{where} \quad V_Q(s) := \max_{a \in \mathcal{A}} Q(s,a).$$

Theorem 1.8 (AJKS): Q = Q* if and only if Q satisfies the Bellman optimality equations.

Value iterations for MDP planning

- The value iteration algorithm iteratively applies the Bellman operator until it converges.
    1. Initialize Q_0 arbitrarily $Q_0 \equiv 0$
    2. for i in 1,2,3,…, k, update $Q_i = \mathcal{T} Q_{i-1}$
    3. Return Q_k

Value iterations for MDP planning

- The value iteration algorithm iteratively applies the Bellman operator until it converges.
    1. Initialize Q_0 arbitrarily
    2. for i in 1,2,3,…, k, update $$\mathcal{T} Q_i = \mathcal{T} Q_{i-1}$$
    3. Return Q_k

1. $\lim_{k\to\infty} Q_k = Q^*$?

- **What is the right question to ask here?**

2. $\|Q_k - Q^*\|_\infty \le \varepsilon(k)$ rate of converge

3. Iteration complexity: ε as an input $k \ge func(\varepsilon)$

Convergence analysis of VI

$\mathcal{T}Q = r + \gamma P \cdot V_Q$, $V_Q = \max_a Q(\cdot, a)$

- Lemma 1. The Bellman operator is a γ-contraction.

$$\textit{For any two vectors } Q, Q' \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|},$$

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \le \boxed{\gamma\|Q - Q'\|_\infty}$$

$\Sigma = 1$

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty = \gamma\|PV_Q - PV_{Q'}\|_\infty = \gamma\|P(V_Q - V_{Q'})\|_\infty$$

operator norm of P in ℓ_∞

$$\to \le \gamma\|V_Q - V_{Q'}\|_\infty = \gamma \max_s |V_Q(s) - V_{Q'}(s)| \le \gamma \max_{s,a} |Q(s,a) - Q'(s,a)|$$

① if $V_Q(s) \ge V_{Q'}(s)$ for those s

$$\gamma \max_s (V_Q(s) - V_{Q'}(s)) \le \gamma \max_a Q(s,a) - \max_a Q'(s,a)$$

$$\le \gamma\, Q(s,a) - Q'(s,a)$$

$$\le \gamma\, |Q(s,a) - Q'(s,a)|$$

$a = \arg\max_a Q(s,a)$

Convergence analysis of VI

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

- Lemma 2. Convergence of the Q function.

$Q' = Q^*$, $\qquad \boxed{TQ^* = Q^*}$

$$\|Q_k - Q^*\|_\infty = \|TQ_{k-1} - Q^*\|_\infty \le \gamma \|Q_{k-1} - Q^*\|_\infty$$

$$\|Q_0 - Q^*\|_\infty \le \frac{1}{1-\gamma} \qquad \cdots \le \gamma^k \cdot \frac{1}{1-\gamma}$$

$$= \frac{(1-(1-\gamma))^k}{1-\gamma} \le \boxed{\frac{e^{-(1-\gamma)k}}{1-\gamma}}$$

$0 \le r(s,a) \le 1 \qquad \left|\sum_{t=1}^{\infty} \gamma^{t-1} r(s,a)\right| \le \frac{1}{1-\gamma}$

Quiz 3: Computing "Iteration complexity" from "convergence bound"?

$$\varepsilon = \frac{e^{-(1-\gamma)k}}{1-\gamma} \Longleftrightarrow \boxed{k = \frac{\log \frac{1}{\varepsilon(1-\gamma)}}{1-\gamma}}$$

$$\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^n = e^{-1}$$

$$\left(1-\frac{1}{n}\right)^n \le e^{-1} \text{ for all } n \ge 1$$

Convergence of the Q function implies the convergence of the value of the *induced* policy.

$\pi_Q^{(s)} = \arg\max_a Q(s,a)$

Lemma 1.11 AJKS (Q-error amplification):

$$V^{\pi_Q} \geq V^\star - \frac{2\|Q - Q^\star\|_\infty}{1-\gamma}\mathbb{1}.$$

Proof: Fix state s and let $a = \pi_Q(s)$. We have:

$$\begin{aligned}
V^\star(s) - V^{\pi_Q}(s) =& Q^\star(s, \pi^\star(s)) - Q^{\pi_Q}(s,a) \\
=& Q^\star(s, \pi^\star(s)) - Q^\star(s,a) + Q^\star(s,a) - Q^{\pi_Q}(s,a) \\
=& Q^\star(s, \pi^\star(s)) - Q^\star(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^\star(s') - V^{\pi_Q}(s')] \\
\leq& \, Q^\star(s, \pi^\star(s)) - Q(s, \pi^\star(s)) + Q(s,a) - Q^\star(s,a) \\
&+ \gamma \mathbb{E}_{s' \sim P(s,a)}[V^\star(s') - V^{\pi_Q}(s')] \\
\leq& \, 2\|Q - Q^\star\|_\infty + \gamma \|V^\star - V^{\pi_Q}\|_\infty.
\end{aligned}$$

where the first inequality uses $Q(s, \pi^\star(s)) \leq Q(s, \pi_Q(s)) = Q(s,a)$ due to the definition of π_Q.

An alternative method: policy iteration

Initialize a policy π_0 arbitrarily.
for k= 1,2,3,4,…

Solution to Bellman equation for π

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

1. *Policy evaluation*. Compute Q^{π_k}
2. *Policy improvement*. Update the policy: $\pi_{k+1} = \pi_{Q^{\pi_k}}$

Theorem 1.14. *(Policy iteration convergence). Let π_0 be any initial policy. For $k \geq \frac{\log \frac{1}{(1-\gamma)\epsilon}}{1-\gamma}$, the k-th policy in policy iteration has the following performance bound:*

$$Q^{\pi^{(k)}} \geq Q^\star - \epsilon \mathbb{1} \, .$$

Computational complexity of these MDP solvers

- VI: $\underbrace{S^2 \cdot A}_{\text{apply } T} \cdot \frac{\log \frac{1}{(1-\gamma)^2 \varepsilon}}{1-\gamma}$ $\quad \varepsilon = 0$
- PI: $(SA)^3 \cdot \frac{\log \frac{1}{(1-\gamma)\varepsilon}}{1-\gamma} \Rightarrow (S^3 + S^2 A) \cdot \frac{\log \frac{1}{(1-\gamma)\varepsilon}}{1-\gamma}$
- LP: $\mathrm{poly}(S, A)$

Strongly polynomial algorithms are independent to ε

a version of simplex method

	Value Iteration	Policy Iteration	LP-Algorithms
Poly?	$\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert\frac{L(P,r,\gamma)\log\frac{1}{1-\gamma}}{1-\gamma}$	$(\lvert\mathcal{S}\rvert^3+\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert)\frac{L(P,r,\gamma)\log\frac{1}{1-\gamma}}{1-\gamma}$	$\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert L(P,r,\gamma)$
Strongly Poly?	✗	$(\lvert\mathcal{S}\rvert^3+\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert)\cdot\min\left\{\frac{\lvert\mathcal{A}\rvert^{\lvert\mathcal{S}\rvert}}{\lvert\mathcal{S}\rvert},\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert\log\frac{\lvert\mathcal{S}\rvert^2}{1-\gamma}}{1-\gamma}\right\}$	$\lvert\mathcal{S}\rvert^4\lvert\mathcal{A}\rvert^4\log\frac{\lvert\mathcal{S}\rvert}{1-\gamma}$

$|S|^2$

Next lecture

- Approximate / randomized solvers for MDP

- MDP / RL with generative models