CS292F StatRL Lecture 2 Markov Decision Processes

Instructor: Yu-Xiang Wang

Spring 2021

UC Santa Barbara

Recap: Markov Decision processes (MDP) parameterization

- Infinite horizon / discounted setting

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

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

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

Initial state distribution $\mu_0 \in \Delta(\mathcal{S})$

Discounting factor: γ

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

$$V^{\pi}(s) = \mathbb{E}_{\pi}[R_1 + \gamma R_2 + ... + \gamma^{t-1} R_t + ... | S_1 = s]$$

- state-action value function: $Q^{\pi}(s,a)$
  - expected **long-term** return when starting in *s*, performing *a*, and following π

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

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

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

- 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

  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

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

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

$$Q^{\pi}(s,a) = ?$$

Deriving Bellman Equation for stationary policies

Bellman equations in matrix forms

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

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

- In matrix forms:

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

Closed-form solution for solving for value functions

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

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

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

Duality between *value functions* and *occupancy measures*

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:

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

Bellman optimality equations characterizes the optimal policy

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

- 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:

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

- Solve for V* by solving the following LP

$$\begin{aligned} \min \quad & \sum_s \mu(s) V(s) \\ \text{subject to} \quad & V(s) \geq r(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s') \quad \forall a \in \mathcal{A},\ s \in \mathcal{S} \end{aligned}$$

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

The Lagrange dual of the LP

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

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

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

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

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

Value iterations for MDP planning

- Recall: Bellman optimality equations

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

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

$$\mathcal{T}Q = r + PV_Q \quad \text{where} \quad V_Q(s) := \max_{a \in \mathcal{A}} Q(s,a).$$

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
  2. for i in 1,2,3,…, k, update $Q_i = \mathcal{T} Q_{i-1}$
  3. Return Q_k
- **What is the right question to ask here?**

Convergence analysis of VI

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

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

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

Convergence analysis of VI

- Lemma 2. Convergence of the Q function.

Quiz 3: Computing “Iteration complexity” from “convergence bound”?

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

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,…

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:
- PI:
- LP:

Strongly polynomial algorithms are independent to ε

	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?	✗	$(\lvert\mathcal{S}\rvert^3+\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert)\cdot\min\left\{\frac{\lvert\mathcal{A}\rvert^{\lvert\mathcal{S}\rvert}}{\lvert\mathcal{S}\rvert},\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert\log\frac{\lvert\mathcal{S}\rvert^2}{1-\gamma}}{1-\gamma}\right\}$	$\lvert\mathcal{S}\rvert^4\lvert\mathcal{A}\rvert^4\log\frac{\lvert\mathcal{S}\rvert}{1-\gamma}$

Next lecture

- Approximate / randomized solvers for MDP

- MDP / RL with generative models