CS292F Statistical Foundation of Reinforcement Learning **Spring 2021**

Lecture 2: Markov Decision Process (Part I), March 31

Lecturer: Yu-Xiang Wang *Scribes: Mengye Liu*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Recap:

Markov Decision processes(MDP) parameteriztion

1. 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. $P(S' \mid S, a)$
   - (Expected) reward function: $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}/[0, R_{\max}]$, $\mathbb{E}\left[R_t \mid S_t = s, A_t = a\right] =: r(s, a)$
     WLOG, we can let $R_{\max} = 1$
   - Innitial state distribution: $\mu_{\cdot} \in \Delta(S)$
   - Discounting factor: $\gamma \in [0, 1]$
     e.g. Horizon $\frac{1}{1-\gamma} = 1 + \gamma + \gamma^2 + \dots$

2. Immediate reward function $r(s, a, s')$

   Expected immediate reward

$$r(s, a, s') = \mathbb{E}[R_1 \mid S_1 = s, A_1 = a, S_2 = s']$$
$$r^\pi(s) = \mathbb{E}_{a\sim\pi(a|s)}[R_1 \mid S_1 = s]$$

3. state value function $V^\pi(s)$

   Expected long-term return when starting in s and following π

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

4. state-action value function $Q^\pi(s, a)$

   Expected long-term return when starting in s, performing a, and following π.

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

5. Optimal value function and the MDP planning problem

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

   Goal of MDP planning is to find π^* such that $V^\pi(s) = V^*(s)$ for all s. For computational reasons, we sometimes want to solve the approximate solution for the problem. We say π is ε- optimal if $V^\pi \geq V^*(s) - \varepsilon\mathbf{1}$.

6. Policies

    - 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

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

7. Few results about MDPs

Proposition It suffices to consider stationary policies.

- Occupancy measure

$$\nu_\mu^\pi(s) = \sum_{t=1}^{\infty} \gamma^{t-1} d^\pi(S_t = s) \quad \text{(State occupancy measure)}$$

$$\nu_\mu^\pi(s,a) = \sum_{t=1}^{\infty} \gamma^{t-1} d^\pi(S_t = s, A_t = a) \quad \text{(State-action occupancy measure)}$$

where $d^\pi(S_t = s)$ is marginal density function under policy π at time t observe state s. Similarly, $d^\pi(S_t = s, A_t = a)$ is marginal distribution policy π at time t with state-action pair (s, a) observed.
Then

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

- There exists a stationary policy with the same occupancy measure.
For a policy π is optimal or any policies π which is non-stationary, $\exists \pi'$ is stationary s.t. $\nu^\pi(s,a) = \nu^{\pi'}(s,a)$.

Corollary There is a stationary poly that is optimal for all initial states.

2.1 Bellman Equations

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

$$V^\pi(s) = \sum_a \pi(a \mid s) \sum_{s'} P(s' \mid s, a) \left[r(s,a,s') + \gamma V^\pi(s') \right] = \sum_a \pi(a \mid s) Q^\pi(s,a) \tag{2.1}$$

Exercise:

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

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

Now we are going to derive Bellman Equation for stationary policies.

$$\begin{aligned}
V^\pi(s) &= \mathbb{E}^\pi\left[\sum_{t=1}^{\infty}\gamma^{t-1}r(S_t,A_t)\mid S_1=s\right]\\
&= \mathbb{E}^\pi\left[r(S_1,A_1)\mid S_1=s\right]+\sum_{S_2}P^\pi(S_2=s'\mid S_1=s)\mathbb{E}^\pi\left[\sum_{t=2}^{\infty}\gamma^{t-1}r(S_t,A_t)\mid S_2=s'\right] \quad \text{Let } \tilde{t}=t-1\\
&= r^\pi(s)+\gamma\sum_{S_2}P^\pi(S_2=s'\mid S_1=s)\mathbb{E}^\pi\left[\sum_{\tilde{t}=1}^{\infty}\gamma^{\tilde{t}-1}r(S_{\tilde{t}},A_{\tilde{t}})\mid S_1=s'\right]\\
\text{By Stationarity } &= r^\pi(s)+\gamma\sum_{S_2}P^\pi(S_2=s'\mid S_1=s)V^\pi(s')
\end{aligned}$$

where $P^\pi(s'\mid s)=\sum_a P(s'\mid s,a)\cdot\pi(a\mid s)$.

We can also write Bellman Equation in matrix form.

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

where $\boldsymbol{P}^\pi\in\mathbb{R}^{S\times S}$ is the transpose of transition matrix under policy π, $\boldsymbol{V}^\pi,\boldsymbol{r}^\pi\in\mathbb{R}^S$.

Lemma 2.1 (Bellman consistency)**.** *For stationary policies, we have*

$$\begin{aligned}
V^\pi &= Q^\pi(s,\pi(s))=\mathbb{E}_{a\sim\pi(a|s)}[Q^\pi(s,a)]\\
Q^\pi(s,a) &= r(s,a)+\gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[V^\pi(s')]
\end{aligned}$$

In matrix forms:

$$\begin{aligned}
\boldsymbol{V}^\pi &= \boldsymbol{r}^\pi+\boldsymbol{\gamma}\boldsymbol{P}^\pi\boldsymbol{V}^\pi \quad \boldsymbol{P}^\pi\in\mathbb{R}^{S\times S}\\
\boldsymbol{Q}^\pi &= \boldsymbol{r}+\boldsymbol{\gamma}\boldsymbol{P}\boldsymbol{V}^\pi\\
\boldsymbol{Q}^\pi &= \boldsymbol{r}+\boldsymbol{\gamma}\boldsymbol{P}^\pi\boldsymbol{Q}^\pi \quad \boldsymbol{P}^\pi\in\mathbb{R}^{SA\times SA}
\end{aligned}$$

where $\boldsymbol{r}\in\mathbb{R}^{SA}$*,* $\boldsymbol{r}^\pi\in\mathbb{R}^S$*.*

Notice: The dimensions of two $\boldsymbol{P}^\pi$'s are different. Both of them are depend on π but in slightly different ways. The first $\boldsymbol{P}^\pi$ is marginal over a and the second $\boldsymbol{P}^\pi$ is joint with a'.

The matrix forms can help us solve the close form of $\boldsymbol{V}^\pi$ and $\boldsymbol{Q}^\pi$. For example, $(\boldsymbol{I}-\boldsymbol{\gamma}\boldsymbol{P}^\pi)\boldsymbol{V}^\pi=\boldsymbol{r}^\pi$, then we can obtain $\boldsymbol{V}^\pi$ by solving this linear equations.

It is interesting that we can connect the matrix forms of value functions with occupancy measure.

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

What we derived in Lecture 1 is that there is also a Bellman equation holds for ν_μ^π.

$$\begin{aligned}
\nu^\pi(s) &= \mu(s)+\gamma\sum_{s'}\nu^\pi(s')P^\pi(s\mid s')\\
\nu^\pi(s,a) &= \mu(s)\pi(s,a)+\gamma\sum_{s'}\nu^\pi(s')\pi(a\mid s)\sum_{a'}P^\pi(s\mid s',a')\pi(a'\mid s')\\
\Rightarrow\nu^\pi(s,a) &= \mu^\pi(s,a)+\gamma\sum_{s'}\sum_{a'}\nu^\pi(a',s')P^\pi(s,a\mid s',a')
\end{aligned}$$

$\begin{cases} V^\pi = (I - \gamma P^\pi)^{-1} r^\pi \\ \nu^\pi = \left(I - \gamma (P^\pi)^T\right)^{-1} \mu \end{cases}$. They are dual to each other in some sense.

To prove that the above equations hold, we need to prove the matrix $I - \gamma P^\pi$ is invertible.

Corollary 2.2. *The matrix $I - \gamma P^\pi$ is full rank/ invertible for any $\gamma < 1$.*

Proof. WTP: $\forall x \neq 0, (I - \gamma P^\pi)x \neq 0$, where I is identity matrix.

$$\begin{aligned} \|(I - \gamma P^\pi)x\|_\infty &= \|x - \gamma P^\pi\|_\infty \\ &\geq \|x\|_\infty - \gamma \|P^\pi x\|_\infty \quad \text{By triangle inequality and linearity} \\ &\geq \|x\|_\infty - \gamma \|x\|_\infty \end{aligned}$$

P^π is a transpose of transition matrix, i.e. each row of P^π is probability distribution ($P(s' \mid s)$), that is the row sum is 1.

By Holder's inequality,

$$P^\pi x = \begin{pmatrix} \langle P^\pi[1, :], x \rangle \\ \vdots \\ \langle P^\pi[n, :], x \rangle \end{pmatrix} = (1 - \gamma)\|x\|_\infty$$

Consider the first element in the vector, $\langle P^\pi[1, :], x \rangle \leq \|P^\pi[1, :]\|_1 \|x\|_\infty \leq \|x\|_\infty$. □

Bellman optimality equations characterizes the optimal policy.

$$V^* = \max_a \sum_{s'} P(s' \mid s, a) \left[r(s, a, s') + \gamma V^*(s') \right] \tag{2.2}$$

where $\sum_{s'} P(s' \mid s, a) r(s, a, s')$ is the expected immediate reward, $\sum_{s'} P(s' \mid s, a) \gamma V^*(s')$ represents discounted future reward by optimal policy.

This is a system of n non-linear equations. If we can solve $V^*(s)$ then it is easy to extract the optimal policy by simply converting it to Q^* function. Then $\pi^*(s) = \operatorname{argmax}_a Q^*(s, a)$.

Proposition 2.3. *There is a deterministic, stationary and optimal policy and it is given by*

$$\pi^*(s) = \operatorname*{argmax}_a Q^*(s, a)$$

Proof. π^* is stationary.

$$\begin{aligned} V^*(s) = V^{\pi^*}(s) &= \mathbb{E}_{a \sim \pi^*(a|s)} \left[Q^{\pi^*}(s, a) \right] \\ &\leq \max_a Q^{\pi^*}(s, a) \\ &= \max_a Q^*(s, a) \quad \text{By the fact } \pi^* \text{ is optimal} \end{aligned}$$

Then define $\pi'(s) = \operatorname{argmax}_s Q^*(s, a)$.

I. Check π' is stationary, i.e. only depends on §.

II. π' is deterministic, i.e.

$$\max_a Q^*(s, a) = Q^*(s, \pi'(s)) \overset{Stationary}{=} V^{\pi'}(s)$$

By definition,

$$V^*(s) \geq V^{\tilde{\pi}}(s), \quad \forall \tilde{\pi}$$

substitute $\tilde{\pi} = \pi'$, then we can get

$$V^{\pi'}(s) \leq V^*(s) \leq V^{\pi'}(s) \Leftrightarrow V^*(s) = V^{\pi'}(s) = \operatorname*{argmax}_a Q^*(s, a)$$

□

2.2 Solving MDP planning problem

The crux of solving a MDP planning problem is to construct Q^*. There are two approaches

- By solving a linear program
- By solving Bellman equations/ Bellman optimality equations

2.2.1 Linear programming approach

Solve for V^* by solving the following LP

$$\begin{aligned} &\min_{V \in \mathbb{R}^S} \sum_s \mu(s) V(s) \\ \text{s.t.} \quad & V(s) \geq \max_a r(s, a) + \gamma \sum_{s'} P(s' \mid s, a) V(s') \quad \forall a \in \mathcal{A}, s \in \mathcal{S} \end{aligned} \tag{2.3}$$

If we substitute $V = V^*$, we have $\sum_s \mu(s) V^*(s) = V^*(\mu)$. The constraints are equivalent to

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

The Lagrange dual of the LP

$$\begin{aligned} &\max_\nu \sum_{s,a} \nu(s, a) r(s, a) \\ \text{s.t.} \quad & \nu \geq 0 \\ & \sum_z \nu(s, a) = \mu(s) + \gamma \sum_{s', a} P(s \mid s', a) \nu(s', a') \end{aligned} \tag{2.4}$$

Linear programming has strong duality, i.e. the minimum of the primal problem is the maximum of the dual problem.

Exercise: Derive the dual by applying the standard procedure.

- Construct Lagrangian multiplier.
- Minimize the Lagrangian to obtain the exact formula

Quiz: Once we have the solution ($\nu \in \mathbb{R}^{SA}$), how to construct the policy?

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

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

When the optimal solution is unique then always exists stationary, deterministic policy.

2.2.2 Value Iteration Algorithm

According to Bellman optimality equations 2.2, we can get

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

Then, we can define

$$\mathcal{T}Q = r + PV_Q$$

where $\mathcal{T}$ is a nonlinear operator, $V_Q(s) := \max_{a \in \mathcal{A}} Q(s,a)$.

Theorem 2.4. *$Q = Q^*$ if and only if Q satisfies the Bellman optimality equations.*

Algorithm: Value iteration(VI)

1. Initialize Q_0 arbitrarily
2. For i in $1, 2, \ldots, k$, update $Q_i = \mathcal{T}Q_{i-1}$
3. Return Q_k

Value iteration algorithm iteratively applies the Bellman operator until it converges.

2.2.3 Convergence analysis of Value Iteration

Lemma 2.5. *The Bellman operator is a γ-contraction. That is $\forall Q, Q' \in \mathbb{R}^{SA}$,*

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

Proof.

$$\begin{aligned}
\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty &= \|r + \gamma PV_Q - (r + \gamma PV_{Q'})\|_\infty \\
&= \gamma \|PV_Q - PV_{Q'}\|_\infty \\
&= \gamma \|P(V_Q - V_{Q'})\|_\infty \quad P \text{ is a linear operator with row sum 1} \\
&\leq \gamma \|V_Q - V_{Q'}\|_\infty \\
&= \gamma \max_s |V_Q(s) - V_{Q''}(s)| \quad \text{By def of } l_\infty \text{ norm}
\end{aligned}$$

1. For s s.t. $V_Q(s) \geq V_{Q'}(s)$, let $a = \text{argmax}_s\, Q(s,a)$

$$\begin{aligned}
\gamma \max_s (V_Q(s) - V_{Q'}(s)) &\leq \gamma Q(s,a) - \max_a Q'(s,a) \\
&\leq \gamma (Q(s,a) - Q'(s,a)) \\
&\leq \gamma |Q(s,a) - Q'(s,a)|
\end{aligned}$$

2. For s s.t. $V_Q(s) < V_{Q'}(s)$, we can get the same conclusion similarly.

Then

$$\gamma \max_s |V_Q(s) - V_{Q'}(s)| \leq \gamma \max_{s,a} |Q(s,a) - Q'(s,a)|$$

□

This lemma shows that the distance of any pairs gets smaller after Bellman operator. Here we set $\gamma < 1$, then the distance tends to zero with exponential rate.

Lemma 2.6. *(Convergence of Q function)*

$$\|Q_k - Q^*\|_\infty \le \frac{e^{-(1-\gamma)k}}{1-\gamma}$$

Proof. Recall that $r(s,a) \in [0,1]$ then $|\sum_{t=1}^{\infty} \gamma^{t-1} r(s,a)| \le \frac{1}{1-\gamma}$ by geometric series. Thus

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

$$\begin{aligned} \|Q_k - Q^*\|_\infty &= \|\mathcal{T}Q_{k-1} - Q^*\|_\infty \\ &\le \gamma\|Q_{k-1} - Q^*\| \quad \text{By lemma 2.5} \\ &\le \cdots \\ &\le \gamma^k \frac{1}{1-\gamma} = \frac{(1-(1-\gamma))^k}{1-\gamma} \\ &\le \frac{e^{-(1-\gamma)k}}{1-\gamma} \end{aligned}$$

The last inequality uses

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

□

Quiz: Compute "iteration complexity" from "convergence bound".

Set $\varepsilon = \frac{e^{-(1-\gamma)k}}{1-\gamma}$, then solve this equation to get

$$k = \frac{\log(\varepsilon(1-\gamma))}{-(1-\gamma)}$$

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

Let $\pi_Q(s) = \operatorname{argmax}_a Q(s,a)$

Lemma 2.7. *($Q-$error amplification)*

$$V^{\pi_Q} \ge V^* - \frac{2\|Q - Q^*\|_\infty}{1-\gamma}\mathbf{1}$$

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

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

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

2.2.4 Policy iteration

An alternative method is policy iteration.

Algorithm: Policy iteration

1. Initialize π_0 arbitrarily
2. For k in $1, 2, \ldots$
   (a) Policy evaluation. Compute Q^{π_k} by solving $Q^\pi = (I - \gamma P^\pi)^{-1} r$.
   (b) Policy improvement. Update the policy: $\pi_{k+1} = \pi_{Q^{\pi_k}}$

Theorem 2.8. *(Policy iteration convergence). Let π_0 be any initial policy. For $k \geq \frac{\log \frac{1}{(1-\gamma)\varepsilon}}{1-\gamma}$, the k-th policy in policy iteration has the following performance bound:*

$$Q^{\pi^{(k)}} \geq Q^* - \varepsilon \mathbf{1}$$

2.2.5 Computational complexity

The computational complexity of three above MDP solvers are as below

Table 2.1: Table of Time Complexity

Value Iteration	Policy Iteration	LP-Algorithm
$S^2 A \cdot \frac{\log \frac{1}{(1-\gamma)^2\varepsilon}}{1-\gamma}$	$(SA)^3 \frac{\log \frac{1}{(1-\gamma)\varepsilon}}{1-\gamma}$	$\text{poly}(S, A)$

For policy iteration, $(SA)^3$ is the time complexity to get the inverse of $(I - \gamma P^\pi)$ naively. It can be improved as $S^3 + S^2 A$. Then the time complexity for PI algorithm will be impoved as $(S^3 + S^2 A) \frac{\log \frac{1}{(1-\gamma)\varepsilon}}{1-\gamma}$.