RL Theory[1]: Lecture 1 (Chapter 1)

Shivam Garg, RLAI@UAlberta

27th August 2020

[1] based on `https://rltheorybook.github.io/`

Adminstrivia (to set the mood right)

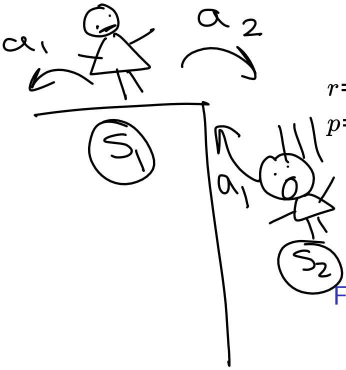
- Insipiration: Sal Khan and Mark Schmidt.
- My goal with these lectures is that **all of you** understand most of the material.
- This partly comes from my own frustration with how inaccessible theory seems to me.
- So I will go slowly and try to be ~~shameless~~ about it. However, I have this problem where I start picking pace without realizing.
- Do stop me at any time if something is unclear (be it my accent, a mistake, some skipped steps).
- And most importantly: “There are 4 kinds of people.”

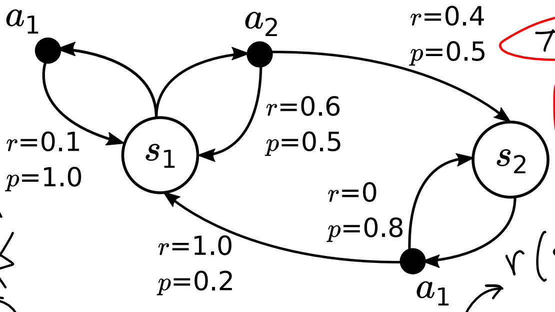
				♡
KL	✓	✗	✗	✓
Theory	✗	✓	✗	✓

Markov Decision Process

$\mu(s_1) = 1$ → formalize world

$\mu(s_2) = 0$





$r(s, A)$

Figure: (**Rooftop**) MDP $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu)$.

$\{s_1, s_2\}$ $\{a_1, a_2\}$ $P(s'|s,a)$

Agent and its Policy

$\pi : S \to \Delta(A)$

$\pi(a_1|s_1) = 1$

$\pi(a_2|s_1) = 0$

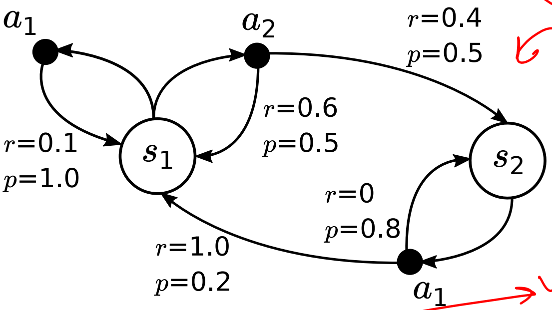
$\pi(a_1|s_2) = 1$

$\pi : S \times A \to [0,1]$

Agent's Interaction and the Trajectory ($\sim$ stream of experience)

$s_1, a_1, 0.1, s_1, a_1, 0.1, \ldots$

M π



$\sim P(\cdot \mid S_0, A_0)$

Figure: Trajectory $\tau_t = (S_0, A_0, R_1, S_1, A_1, R_2, S_2, \ldots, S_t)$.

$\sim M(\cdot)$ $\sim \pi(\cdot \mid S_0)$ $r(S_0, A_0)$

Goal of the Agent

safe $\quad \pi$

$R \in [0,1]$

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

- Agent interacts with the environment to generate a trajectory

$$\tau_t = (S_0, A_0, R_1, S_1, A_1, R_2, S_2, \ldots, S_t).$$

$\|Q^* - Q\|_\infty \leq \frac{1}{(1-\gamma)}$

- Then agent learns to optimize its policy π to maximize the return (in expectation)

$$[0,1) \qquad G_0 = \overset{(1-\gamma)}{\cancel{\frac{1}{1-\gamma}}}(\overset{1}{R_1} + \gamma \overset{1}{R_2} + \gamma^2 \overset{1}{R_3} + \cdots).$$

safe π

- Example: For the trajectory $\tau = (s_1, a_1, 0.1, s_1, a_1, 0.1, \ldots)$ and $\gamma = 0.9$,

$$G_0 = (1-\gamma)\left[0.1 + \gamma\, 0.1 + \gamma^2\, 0.1 + \cdots\right]$$

$$= \boxed{0.1}$$

State Value function

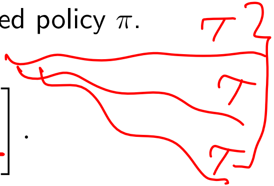
$\mu \qquad \pi \qquad P \qquad\qquad \mu \;\; \pi \;\; P$

- On what factors is the trajectory dependent?

$$\tau_t = (S_0, \underline{A_0}, R_1, \underline{S_1}, A_1, R_2, S_2, \ldots, S_t).$$

- Value function allows the agent to predict the expected return (under the environment's transition dynamics) from a given state for a fixed policy π.
- State Value function $V^\pi : \mathcal{S} \to \mathbb{R}$

$$V^\pi(s) = \boxed{(1-\gamma)}\, \mathbb{E}_{P,\pi}\left[\sum_{t=0}^{\infty} \gamma^t \underbrace{r(S_t, A_t)} \middle| S_0 = \underline{s}\right].$$



- Expectation $\to$ summation notation:

$$V^\pi(s) = \sum_{a_0} \pi(a_0|s)\left[(1-\gamma)\, r(s,a_0) + \sum_{s_1} P(s_1|s,a_0)\left[\sum_{a_1} \pi(a_1|s_1) \ldots\right.\right.$$

Policy Evaluation: V^π for Rooftop MDP

$$(1-\gamma)\, \mathbb{E}_{\pi,P}\left[\sum \gamma^t R \;\middle|\; s_0 = s \right]$$

$s_1, a_1, 0.1, s_1, a_1, 0.1, \ldots$

$\boxed{0.1}$

- Consider the safe policy: $\pi(A|s_1) = \begin{cases} 1 & A = a_1, \\ 0 & A = a_2. \end{cases}$

  and $\pi(a_2|s_2) = 1$.

  π

- Then $V^\pi(s_1) = 0.1$
- What about $V^\pi(s_2) = ?$

$P(\tau)$	τ
0.2	$s_2, a_1, 1, s_1, a_1, 0.1, \ldots$
0.8×0.2	$s_2, a_1, 0, s_2, a_1, 0.1, s_1, \ldots$
$0.8^2 \times 0.2$	$s_2 \quad s_2 \quad s_1$
$\vdots$	$\vdots$

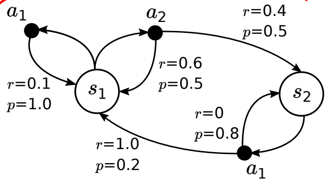


Figure: (**Rooftop**) MDP.

(Policy evaluation is tedious! We'll later describe how to do it iteratively.)

Goal of RL Agent (again!)

s_0

$$\pi^* = \max_{\pi \in \Pi} V^\pi(s) \quad \text{for given state } s$$

π_1 — $V^{\pi_1}(s)$

π_2 — $V^{\pi_2}(s)$

π^* —

State–Action Value Function

S

$a \sim \pi$

S

$$Q^{\pi}(s,a) = \boxed{(1-\gamma)}\,\mathbb{E}_{P,\pi}\left[\sum_{t=0}^{\infty} \gamma^{t} r(s_t,a_t) \,\middle|\, S_0 = s, A_0 = a\right].$$

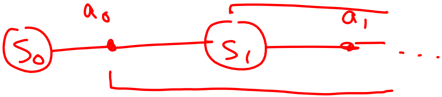
given

a

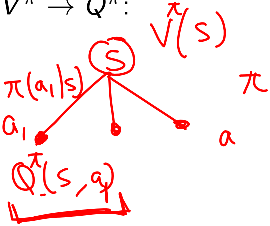
S

π

Bellman (Constitency) Equations



- $V^\pi \to Q^\pi$: $V^\pi(s) = Q^\pi(s, \pi(s))$ for deterministic policy π.

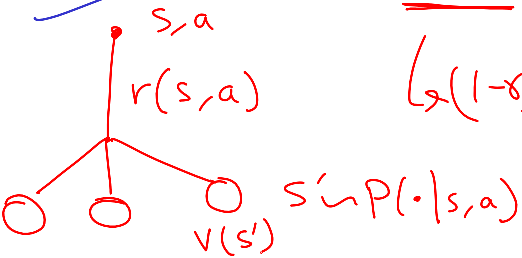


$$V^\pi(s) = \sum_a \underbrace{\pi(a|s)}_{[0,1]} Q^\pi(s,a)$$

normalized

- ~~$Q^\pi \to V^\pi$:~~ $Q^\pi(s, a) = (1 - \gamma) r(s, a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^\pi(s')]$.

$$= (1-\gamma) r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^\pi(s')$$



Reminder about Matrix–Vector Multiplication

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \times 1 + 2 \times 2 \\ 3 \times 1 + 4 \times 2 \end{bmatrix}$$

$$= \begin{bmatrix} 1 \\ 3 \end{bmatrix} \times 1 + \begin{bmatrix} 2 \\ 4 \end{bmatrix} \times 2 \quad \checkmark$$

$$\begin{bmatrix} \begin{pmatrix} \\ \\ \end{pmatrix} & \begin{pmatrix} \\ \\ \end{pmatrix} & \begin{pmatrix} \\ \\ \end{pmatrix} \end{bmatrix}\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{pmatrix} \\ \\ \end{pmatrix} 0 + \begin{pmatrix} \\ \\ \end{pmatrix} 0 + \begin{pmatrix} \\ \\ \end{pmatrix} 0$$

Vector Notation for r, V, and Q

- The reward vector $r \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$:

$$r = \begin{bmatrix} r(s_1, a_1) \\ r(s_1, a_2) \\ r(s_2, a_1). \end{bmatrix} = \begin{bmatrix} 0.1 \\ 0.5 \\ 0.2 \end{bmatrix}$$

$Q(s, a)$

- Value function vectors $V^\pi \in \mathbb{R}^{|\mathcal{S}|}$ and $Q^\pi \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$:

$V(s)$

$$V^\pi = \begin{bmatrix} V^\pi(s_1) \\ V^\pi(s_2) \end{bmatrix}; \qquad Q^\pi = \begin{bmatrix} Q^\pi(s_1, a_1) \\ Q^\pi(s_1, a_2) \\ Q^\pi(s_2, a_1) \end{bmatrix}.$$

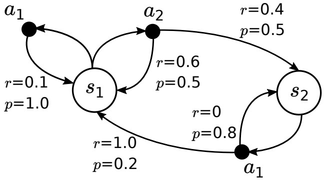


Figure: (**Rooftop**) MDP.

Vector Notation for P and P^π

P $\quad$ P^π

- $P \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$, with entry $P_{(s,a),s'} = P(s'|s,a)$.

S

$$(s,a) \quad \begin{bmatrix} P(s_1|s_1,a_1) & P(s_2|s_1,a_1) \\ P(s_1|s_1,a_2) & P(s_2|s_1,a_2) \\ P(s_1|s_2,a_1) & P(s_2|s_2,a_1) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0.5 & 0.5 \\ 0.2 & 0.8 \end{bmatrix}$$

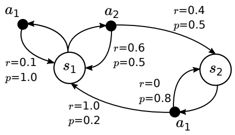


Figure: **(Rooftop)** MDP.

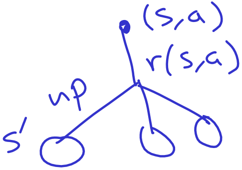
- $P^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$, with entry $P_{(s,a),(s',a')} = P(s'|s,a) \cdot \pi(a'|s')$.

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

$$\begin{bmatrix} s_1a_1|s_1a_1 & s_1a_2|s_1a_1 & s_2a_1|s_1a_1 \\ \vdots & \vdots & \vdots \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ & & \end{bmatrix}$$

3×3

Vector Notation for $Q^\pi \to V^\pi$ Bellman Equation



$s' \mid s,a$

$$\underbrace{Q^\pi}_{3\times1} = (1-\gamma)\underbrace{r}_{3\times1} + \gamma \underbrace{P}_{3\times2}\underbrace{V^\pi}_{2\times1}.$$

$$Q^\pi(s,a) = (1-\gamma)\, r(s,a) + \gamma \sum_{s'} p(s'|s,a)\, V^\pi(s')$$

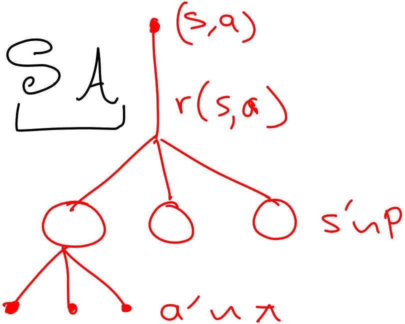
$$\begin{bmatrix} Q^\pi(s_1,a_1) \\ Q^\pi(s_1,a_2) \\ Q^\pi(s_2,a_1) \end{bmatrix} = (1-\gamma)\begin{bmatrix} r(s_1,a_1) \\ \cdot \\ \cdot \end{bmatrix} + \gamma \begin{bmatrix} p(s_1|s_1,a_1) & p(s_2|s_1,a_1) \\ p(s_1|s_1,a_2) & \\ p(s_1|s_2,a_1) & \end{bmatrix} \begin{bmatrix} V(s_1) \\ V(s_2) \end{bmatrix}$$

Vector Notation for $Q^\pi \to Q^\pi$ Bellman Equation

$$\begin{bmatrix} p(s_1 a_1 | s_1 a_1) \\ p(s_1 a_2 | s_1 a_1) \\ p(s_2 a_1 | s_1 a_1) \end{bmatrix}$$

SA

$$Q^\pi = (1-\gamma) r + \gamma P^\pi Q^\pi .$$



$$Q^\pi(s,a) = (1-\gamma) r(s,a) + \gamma \sum_{s'} p(s'|s,a) \sum_{a'} \pi(a'|s') Q^\pi(s',a')$$

$$= (1-\gamma) r(s,a) + \gamma \sum_{s'} \sum_{a'} p(s'|s,a) \pi(a'|s') Q(s',a')$$

$$Q^\pi(s,a) = (1-\gamma) r(s,a) + \gamma \sum_{s',a'} P^\pi(s'a'|sa) Q^\pi(s,a)$$

Exact Solution for Q^π

$$Q^\pi = (1-\gamma)r + \gamma P^\pi Q^\pi.$$

$$Q^\pi - \gamma P^\pi Q^\pi = (1-\gamma) r$$

$$(I - \gamma P^\pi) Q^\pi = (1-\gamma) r$$

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

Solving Exactly for Q^π (Rooftop MDP with Safe Policy)

? $(1-\gamma)(0.1+\gamma 0.1+\gamma^2 0.1+\cdots$

- Calculation:

P^π $\quad r$

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

$$= (1-0.9)\left(\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix} - 0.9\begin{bmatrix}1 & 0 & 0\\ 0.5 & 0 & 0.5\\ 1 & 0 & 0\end{bmatrix}\right)^{-1}\begin{bmatrix}0.1\\ 0.5\\ 0.2\end{bmatrix}$$

$$= \begin{bmatrix}Q^\pi(s_1,a_1)\\ Q^\pi(s_1,a_2)\\ Q^\pi(s_2,a_1)\end{bmatrix} = \begin{bmatrix}0.1\\ 0.1445\\ 0.11\end{bmatrix}$$

$\pi^Q(s)$

s_1 — a_2

s_2 — a_1

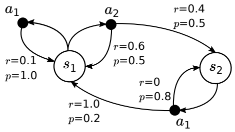


Figure: (**Rooftop**) MDP.

- Define $\pi_Q(s) := \arg\max_{a\in\mathcal{A}} Q(s,a)$.
- Define $V_Q(s) := \max_{a\in\mathcal{A}} Q(s,a)$.

$TQ = (1-\gamma)\begin{bmatrix}\cdot\\ \vdots\end{bmatrix} + \gamma P\begin{bmatrix}0.1445\\ 0.11\end{bmatrix}$

$V_Q(s_1) = 0.1445$ $\quad V_Q(s_2) = 0.11$

$\underbrace{I - \gamma P^\pi}$ is Invertible

$\gamma \in [0, 1)$

- Invertibility = Full Rank

$$\underset{A\ B\ C}{\begin{bmatrix} 0 & 0 & 0 \end{bmatrix}} \begin{array}{c} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \end{array}\begin{array}{c} a \\ b \\ c \end{array} = A\underline{a} + B\underline{b} + C\underline{c} \neq \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

$a, b, c \neq 0$

$$X x \neq 0 \quad \forall x \neq 0$$

- ∞–norm and triangle inequality

$$\|X\|_\infty = \max_x |x|$$

$$\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \rightarrow 3$$

$$\|A\|_\infty + \|B\|_\infty \geq \|A + B\|_\infty$$

$$\|A - B\| + \|B\| \geq \|A\|$$

$$\|A - B\| \geq \|A\| - \|B\|$$

$(I - \gamma P^\pi)$ is Invertible (Proof)

max $|x|$

$x \neq 0$

$$\|\underline{(I - \gamma P^\pi) x}\|_\infty = \|x - \gamma P^\pi x\|_\infty$$

$$\geq \|x\|_\infty - \gamma \|P^\pi x\|_\infty$$

$$\geq \|x\|_\infty - \gamma \|x\|_\infty = (1-\gamma)\|x\|_\infty$$

$$> 0$$

$\gamma = 1 \to 0$

$+\|P_\pi x\|_\infty \leq \|x\|_\infty$

$-\|P_\pi x\|_\infty \geq -\|x\|_\infty$ $\quad x^* = \max_i x_i = \|x\|_\infty$ $\quad \leq P_1 x^* + P_2 x^* + P_3 x^*$

$\sum P_i = 1$ $\quad = x^*$

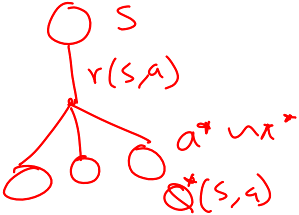
$$\begin{bmatrix} P_1 & P_2 & P_3 \\ & & \\ & & \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} P_1 x_1 + P_2 x_2 + P_3 x_3 \\ \\ \end{bmatrix} \leq \begin{bmatrix} \max x_i \\ \\ \end{bmatrix}$$

Bellman Optimality Equations

- $V^* \to Q^*$

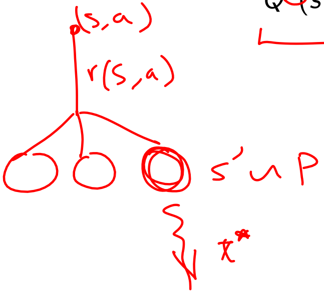
$$V^*(s) = \max_{a \in \mathcal{A}} Q^*(s, a).$$

$$\boxed{V^*(s) = Q^*(s, \pi^*(s))}$$



- $Q^* \to V^*$

$$Q^*(s, a) = (1 - \gamma) r(s, a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')].$$



Bellman Optimality Operator $\mathcal{T}$

- $\mathcal{T} : \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$

$Q^*(s_1, a_1)$
$Q^*(s_1, a_2)$
$Q^*(s_2, a_1)$

$Q^*(s_1, \pi^*(s_1))$

$V^*(s_1)$

$$\mathcal{T}Q := (1-\gamma)r + \gamma P V_Q.$$

$$\mathcal{T}Q(s,a) = (1-\gamma)\, r(s,a) + \gamma \sum_{s'} p(s'|s,a)\, V_Q(s')$$
$$= (1-\gamma)\, r(s,a) + \gamma \sum_{s'} p(s'|s,a) \max_{a'} Q(s',a')$$

$$\mathcal{T}Q(s,a) = (1-\gamma)\, r(s,a) + \max_{a'} \gamma \sum_{s'} p(s'|s,a)\, Q(s', a')$$

- Therefore, the $Q^* \to V^*$ equation can be written as

$$Q^*(s,a) = (1-\gamma)r(s,a) + \gamma\,\mathbb{E}_{s'\sim P(\cdot|s,a)}[V^*(s')]$$

$$\Rightarrow \quad Q^* = \mathcal{T}Q^*.$$

fixed point

$V_{Q^*}(s')$

$$(1-\gamma)r(s,a) + \gamma \sum_{s'} p(s'|s,a)\, V_{Q^*}(s')$$

Q^* $\quad \boxed{V_{Q^*} = V^*}$

Q–Value Iteration

Q^*

iterative eqn. in vectors

- Q–value iteration:

$$\boxed{Q^{(k+1)} = \mathcal{T} Q^{(k)}}.$$

$$Q^{(k)} \xrightarrow{\mathcal{T}} Q^{(k+1)} \xrightarrow{\mathcal{T}} Q^{(k+2)} \quad \cdots$$

$$Q^{(0)} \longrightarrow Q^{(1)} \longrightarrow \cdots \longrightarrow Q^{*}$$

- What are iterative equations?
- E.g.: $x_{k+1} = 4x_k + 3$. (divergent for $x_0 = 0$) $\quad 0, 3, 15, \ldots \longrightarrow$
- E.g.: $y_{k+1} = 0.1 y_k + 1$ (convergent for $y_0 = 0$) $\quad 0, 1, 1.1, 1.11, \ldots$

0.111111

$1 + 0.1 \times 0.11111 = 0.11111 + 1 = 1.11111$

Q–Value Iteration Converges!

$V^* - V^k \leq 0.1$ ✓

Theorem (Q–Value Iteration Convergence)

- Set $Q^{(0)} = 0$.
- Obtain $Q^{(k+1)} = \mathcal{T} Q^{(k)}$ for $k = 0, 1, 2, \ldots$.
- Let $\pi^{(k)} = \pi_{Q^{(k)}}$. $\pi^{(k)}$ is the greedy wrt. $Q^{(k)}$
- Then for $k > \frac{1}{1-\gamma} \log\left(\frac{2}{\epsilon(1-\gamma)}\right)$,

$V^* > V^{\pi^{(k)}}$

$$V^{\pi^{(k)}} \geq V^* - \epsilon \mathbb{1}. \equiv \boxed{V^*(s) - V^{\pi^{(k)}}(s) \leq \epsilon} \; \forall s$$

$V^* - V^{\pi^{(k)}} \leq \epsilon \mathbb{1}$

$O(\log(1/\epsilon))$

$\epsilon = 0.1$ $k = \log(10)$

$\epsilon = 0.01$ $k = \log 100$

Proof: Q–Value Iteration Convergence

- To prove: For $k > \frac{1}{1-\gamma} \log\left(\frac{2}{\epsilon(1-\gamma)}\right)$, $V^{\pi^{(k)}} \geq V^* - \epsilon\mathbb{1}$ holds.

- Bellman Optimality Operator $\mathcal{T}$ is a Contraction

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

- Bounds on $\|Q^{(k)} - Q^*\|_\infty$

$$\|Q^{(k)} - Q^*\|_\infty \leq e^{-(1-\gamma)k}.$$

- Bounding the Suboptimality of π_Q

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

(Part 1): Bellman Optimality Operator $\mathcal{T}$ is a Contraction

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

(Part 2): Bounds on $\|Q^{(k)} - Q^*\|_\infty$

.

$$\|Q^{(k)} - Q^*\|_\infty \leq e^{-(1-\gamma)k}.$$

(Part 3): Bounding the Suboptimality of π_Q

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

(Final Part): Proof of Q–Value Iteration Convergence

Summary

- What is MDP?
- What is an agent?
- The goal of RL.
- Value iterations.
- Next time (play with Rooftop MDP, existence of an optimal stationary and deterministic policy, policy iteration)