MDP Basis

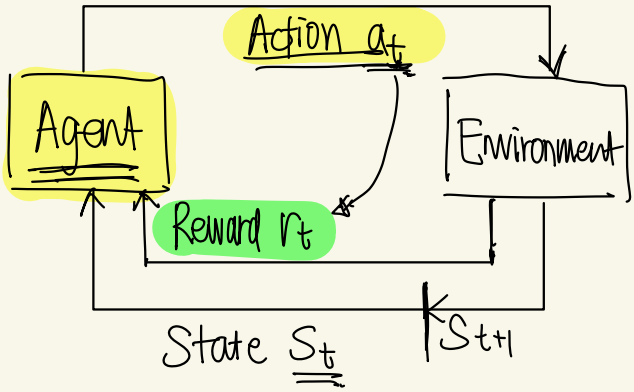
Markov Decision Processes (MDPs)

$$\mathcal{M} = (S, A, P, R, \gamma, d_0)$$

(infinite-horizon discounted MDP)

- S: state space 状态
- A: action space 动作
- $P: S \times A \to \Delta(S)$: transition dynamics 转移 $P(s' \mid s, a)$ simplex
- $R: S \times A \to [0, R_{max}]$: reward 奖赏(收益) $R(s, a)$
- $\gamma \in [0, 1)$: discount factor 折扣因子.
- $d_0 \in \Delta(S)$: initial state distribution 初始状态分布

Interaction



"online"

trajectory: $s_1, a_1, r_1, s_2, a_2, r_2, \cdots\cdots = \tau_{\pi}$

d_0 $\pi(s_1)$ $R(s_1,a_1)$ $P(\cdot|s_1,a_1)$

$$r_i \sim R(s_i, a_i)$$

$$s_{i+1} \sim P(\cdot|s_i, a_i)$$

未来一轮 当前轮

Markovian 马尔可夫性

- d_0 ✓
- π ✓
- R
- P

$X_0, X_1, X_2, \cdots$

Goal: take actions that maximize return 回报

$$\mathbb{E}_{\tau_\pi}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t\right] = J(\pi)$$

$$= \mathbb{E}[r_1 + \gamma \cdot r_2 + \gamma^2 r_3 + \cdots]$$

随机性 (randomness)

Policy 策略

def A (determistic & stationary) policy

$\pi: S \to A$: $a_t = \pi(s_t)$

More generally, stochastic policy

$\pi: S \to \Delta(A)$ $a_t \sim \pi(s_t)$

non-stationary policy

$$\pi : \mathbb{N} \times S \to \Delta(A)$$

$$\pi = \{\pi_t\}_{t=1,2,\dots} \qquad a_t \sim \pi_t(s_t)$$

Value Function

$a_t \sim \pi(\cdot|s_t)$
$r_t \sim R(s_t, a_t)$
$s_{t+1} \sim P(\cdot|s_t, a_t)$

Sutton & Barto

condition expectation.

fixed, s, π

def

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \,\middle|\, s_1 = s, \pi\right]$$

follow this policy

$$\boxed{J(\pi) = \mathbb{E}_{s\sim d_0}\left[V^{\pi}(s)\right]}$$

How to compute V^{π}?

policy evaluation : π

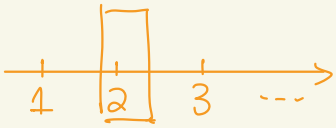
Bellman Equation

(for policy π) $\sum_{a\in A} \pi(a|s) \cdot R(s,a)$

$$V^{\pi}(s) = R(s, \pi(s)) + \gamma \cdot \sum_{s'\in S} P(s'|s, \pi(s))\, V^{\pi}(s')$$

current state

next state



$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \,\middle|\, s_1 = s\right]$$

$$= \mathbb{E}_{\pi}\left[r_1 + \gamma \sum_{t=2}^{\infty} \gamma^{t-2} r_t \,\middle|\, s_1 = s\right]$$

linearity

$$= \mathbb{E}_{\pi}\left[r_1 \,\middle|\, s_1 = s\right] + \gamma \cdot \mathbb{E}_{\pi}\left[\sum_{t=2}^{\infty} \gamma^{t-2} r_t \,\middle|\, s_1 = s\right]$$

$$= R(s, \pi(s)) + \gamma \cdot \mathbb{E}_{\substack{\pi \\ s' \mid s}}\Big[\underbrace{\mathbb{E}_{\pi}\left[\sum_{t=2}^{\infty} \gamma^{t-2} r_t \,\middle|\, s_1 = s, s_2 = s'\right]}_{f(s,s')}\Big]$$

Law of Total Expectation:

$$\mathbb{E}_{Y}\left[\mathbb{E}_{X}[X \mid Y]\right] = \mathbb{E}[X]$$

$$\sum_{s'} P(s' \mid s, \pi(s)) \cdot \mathbb{E}_{\pi}\left[\sum_{t=2}^{\infty} \gamma^{t-2} r_t \,\middle|\, \boxed{s_2 = s'}\right]$$

s_1

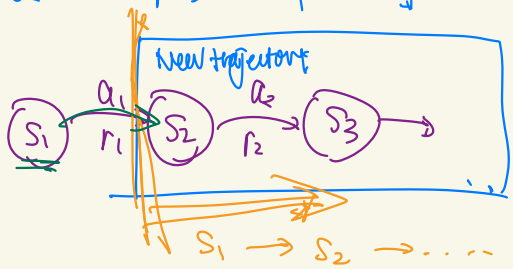
$$= \sum_{s'} P(s' \mid s, \pi(s)) \cdot \mathbb{E}_{\pi}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_{t+1} \,\middle|\, s_2 = s'\right]$$

$$= \sum_{s'} P(s' \mid s, \pi(s)) \cdot \mathbb{E}_{\pi}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \,\middle|\, s_1 = s'\right]$$

$$\mathbb{E}_{\pi}\left[r_2 + \gamma \cdot r_3 + \gamma^2 r_4 + \cdots \,\middle|\, s_2 = s'\right]$$

$$\Rightarrow \mathbb{E}_{\pi}\left[r_1 + \gamma \cdot r_2 + \gamma^2 r_3 + \cdots \,\middle|\, s_1 = s'\right]$$

$$= \sum_{s'} P(s' \mid s, \pi(s)) \cdot V^{\pi}(s')$$



Bellman equation

$$\Rightarrow \boxed{V^{\pi}(s)} = R(s, \pi(s)) + \gamma \cdot \sum_{s'} P(s' \mid s, \pi(s)) \boxed{V^{\pi}(s')}$$

$$\triangleq R(s,\pi(s)) + \gamma \cdot \langle P(\cdot|s,\pi(s)), V^\pi(\cdot)\rangle$$

$$= R(s,\pi(s)) + \gamma \cdot \mathbb{E}_{s'\sim P(\cdot|s,\pi(s))}\left[V^\pi(s')\right] \quad \star$$

Claim: Value Function V^π is the ONLY solution of the Bellman equation (for policy π) $\star\star$

$\forall s$: $V^\pi(s) = R(s,\pi(s)) + \gamma \cdot \langle P(\cdot|s,\pi(s)), V^\pi(\cdot)\rangle$

Matrix form: V^π: $|S| \times 1$ vector $[V^\pi(s)]_{s\in S}$

to be evaluate

R^π: $|S| \times 1$ vector $[R^\pi(s,\pi(s))]_{s\in S}$

known

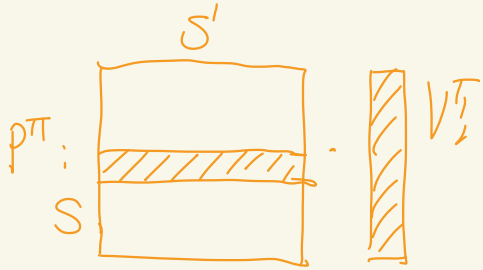
P^π: $|S| \times |S|$ matrix $[P^\pi]_{s,s'} = P(s'|s,\pi(s))$

known

$$V^\pi(s) = R^\pi(s) + \gamma \cdot P^\pi V^\pi(s)$$

$$\Downarrow$$

$$V^\pi = R^\pi + \gamma \cdot P^\pi V^\pi$$



$$\Rightarrow (I-\gamma\cdot P^\pi)\cdot V^\pi = R^\pi \qquad Ax=b$$

invertible

$$\Rightarrow \boxed{V^\pi = (I-\gamma\cdot P^\pi)^{-1} R^\pi}$$

unique solution

矩阵A: A 可逆

$\Leftrightarrow \forall x\neq 0.\ Ax\neq 0 \rightarrow \|Ax\|_\infty > 0$

∞-norm: $\|x\|_\infty = \max_i |x_i|$

$x\in\mathbb{R}^{|S|}$, $x\neq 0$: $\|(I-\gamma\cdot P^\pi)\cdot x\|_\infty = \|x-\gamma\cdot P^\pi x\|_\infty$

$\geq \|x\|_\infty - \gamma\|P^\pi x\|_\infty$ (triangular inequality)

discount factor

$\forall s$: $\left|[P^\pi x]_s\right| = \left|\langle P^\pi(\cdot|s,\pi(s)), x\rangle\right| \leq \|x\|_\infty$

since $\left|\mathbb{E}_{s'\sim P^\pi(\cdot|s,\pi(s))}[x(s')]\right| \leq \max_{s'}|x(s')|$

$(|\mathbb{E}[X]| \leq \mathbb{E}[|X|])$

$\|P^\pi x\|_\infty \leq \|x\|_\infty$

$\Rightarrow \gamma\cdot\|P^\pi x\|_\infty < \|x\|_\infty \ (\gamma\in[0,1]) \Rightarrow \|(I-\gamma\cdot P^\pi)\cdot x\|_\infty > 0.$

Recall: MDP $\mathcal{M} = (S, A, P, R, \gamma, d_0)$

↳ infinite-horizon, discounted

☆ finite-horizon, undiscounted $\sum_{h=1}^{H} r_h$

infinite-horizon average-reward. $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} r_t$

episodic

for $t = 1, 2, \cdots$ do:

☆ for $h = 1, 2, \cdots, H$ do:

$s_{t,h}$, $a_{t,h}$ $r_{t,h}$,

$s_{t,H+1}$: absorbing state.

s_{abs} $r=0$

$\tilde{M} = (\tilde{S}, A, \tilde{P}, \tilde{R}, \gamma, d_0)$

$\tilde{S} = S \times [H] \cup \{s_{absorbing}\}$

$\tilde{P}((s', h+1) \mid (s, h), a) = P(s' \mid s, a)$

$\tilde{R}((s, h), a) = R(s, a)$

$\gamma = 1.$

(* inhomogeneous

不同层的: $\tilde{P}((s', h+1) \mid (s, h), a) = \boxed{P_h}(s' \mid s, a)$)

⇒ Optimal policy: non-stationary $\pi = \{\pi_h\}_{h \in [H]}$

$R: S \times A \to [0, R_{max}]$. wlog.

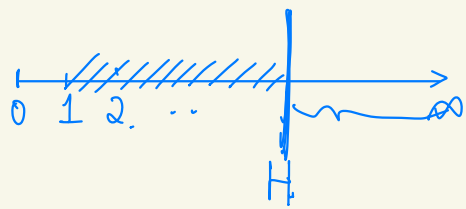
$$\Rightarrow V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \,\middle|\, s_1 = s \right]$$

$$\leq R_{max} \cdot \sum_{t=1}^{\infty} \gamma^{t-1} = \boxed{\frac{R_{max}}{1-\gamma} \triangleq V_{max}} \quad \star$$

$r_t \in [0, R_{max}]$

$\Rightarrow V^{\pi}(s) \in [0, V_{max}]$. $\frac{1}{1-\gamma}$

Infinite-horizon discounted.



$H = \frac{1}{1-\gamma}$

effective horizon: 时域

reward: $R \in \mathbb{R}^{|S \times A|}$ (s,a)

$\Updownarrow$

$R + C \cdot I_{|S \times A|}$

"background noise" $\left(\gamma^{H+1} + \gamma^{H+2} + \cdots\right) R_{max}$

$= \frac{1}{1-\gamma^{H}} = \frac{1}{1-\gamma^{\frac{1}{1-\gamma}}}$

e.g. $R(s,a) \in [-x, y]$. $x, y > 0$.

Offset $C = x$.

$R_{max} = x + y$.

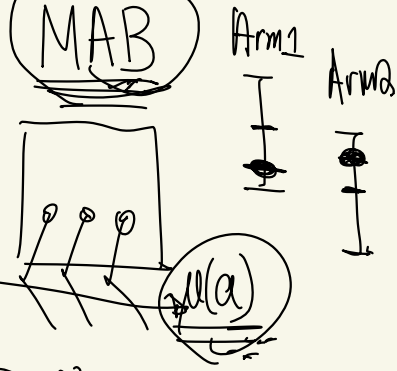
$\Rightarrow R(s,a) \in [0, R_{max}]$

deterministic / stochastic ← bandits.

$R(s,a) = \mathbb{E}[r_t \mid s_t = s, a_t = a]$

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

MAB



$\Rightarrow$ 学习 reward 并不比学习 transition 困难.

仍会影响 learning efficiency : statistical efficiency

Bellman equation (for policy π):

$\pi: S \to A$

$$V^\pi(s) = R(s, \pi(s)) + \gamma \cdot \mathbb{E}_{s' \sim P(\cdot \mid s, \pi(s))}[V^\pi(s')]$$

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

$$V^\pi(s) = \mathbb{E}_{a \sim \pi(s)}[R(s,a)] + \gamma \cdot \mathbb{E}_{\substack{a \sim \pi(s) \\ s' \sim P(\cdot \mid s, \pi(s))}}[V^\pi(s')]$$

Matrix form:

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

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

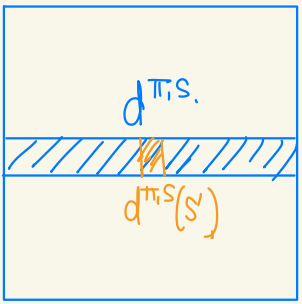
$R^\pi(s) = \mathbb{E}_{a \sim \pi(s)}[R(s,a)]$ $\quad R(s, \pi)$

$P^\pi(s, s') = \mathbb{E}_{a \sim \pi(s)}[P(s' \mid s, a)]$ $\quad P(s' \mid s, \pi)$

State Occupancy

$$\boxed{(1-\gamma)\cdot\left(I-\gamma P^{\pi}\right)^{-1}} \quad |S|\times|S| \text{ matrix}$$

- each row (indexed by s) is the normalized discounted state occupancy $d^{\pi,s} \in \mathbb{R}^{|S|}$.



$$d^{\pi,s}(s') = (1-\gamma)\cdot\sum_{t=1}^{\infty}\gamma^{t-1}\cdot d_t^{\pi,s}(s')$$

discounted average

where $d_t^{\pi,s}(s') = \mathbb{P}^{\pi}\left[S_t = s' \mid S_1 = s\right]$

$$s \xrightarrow{1} s_2 \xrightarrow{2} s_3 \cdots$$

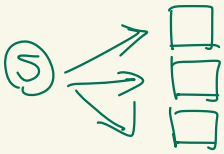
$$V^{\pi}(s) = \sum_{t=1}^{\infty}\gamma^{t-1}\cdot\mathbb{E}_{\pi}\left[r_t \mid S_1 = s\right]$$

State occupancy

Visitation frequency.

$$= \sum_{t=1}^{\infty}\gamma^{t-1}\cdot\sum_{s'}\mathbb{P}^{\pi}\left[S_t = s' \mid S_1 = s\right]\cdot R(s',\pi)$$

$$= \left\langle \underbrace{\sum_{t=1}^{\infty}\gamma^{t-1}\cdot\mathbb{P}^{\pi}\left[S_t = \cdot \mid S_1 = s\right]}_{\text{discounted occupancy}}, \underbrace{R(\cdot,\pi)}_{\text{reward}} \right\rangle$$



$$d^{\pi,s}(s')\cdot R(s',\pi)$$

$$(I-\gamma P^\pi)^{-1} = \sum_{t=0}^{\infty} (\gamma P^\pi)^t = \sum_{t=1}^{\infty} \gamma^{t-1} (P^\pi)^{t-1}$$

transition matrix

- $(1-\gamma)$: normalizing factor.

$\downarrow$

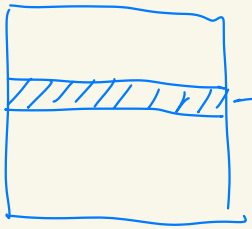
matrix: row-stochastic

$$(P^\pi)^{t-1}(s, s') = \boxed{P^\pi[s_t = s' \mid s_1 = s]}$$

$(t-1)$ transitions

$\Downarrow$

$$V^\pi(s') = (I-\gamma P^\pi)^{-1} \cdot R^\pi$$



→ 和为 1

(row-stochastic)

Optimality

$\pi \to V^\pi$

Motivation: policy-evaluation 策略评估

→ policy-optimization 策略优化

Claim: for infinite-horizon discounted MDPs, there always exists a stationary and deterministic ☆ policy π^* that is optimal for all starting states simultaneously.

$$\forall s \in S \ \forall \pi: \ V^{\pi^*}(s) \geq V^\pi(s)$$

$\Downarrow$

AJKS Chapter 1.3

Bellman Optimality Equation

① $$V^*(s) = \max_{a \in A} \left(R(s,a) + \gamma \cdot \mathbb{E}_{s' \sim p(\cdot|s,a)} [V^*(s')] \right)$$

Claim: there can be many optimal policies,

but the optimal value is unique.

(which is the solution of BOE)

$V^\pi \Leftrightarrow$ BE: $\boxed{V^\pi(s) = R(s,\pi) + \gamma \cdot \mathbb{E}_{s' \sim p(\cdot|s,\pi)} [V^\pi(s')]}$

$V^* \Leftrightarrow$ BOE: $V^*(s) = \max_{a \in A} \left(R(s,a) + \gamma \cdot \mathbb{E}_{s' \sim p(\cdot|s,a)} [V^*(s')] \right)$

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

① How to get V^*?

non-linear

a set of non-linear equations!

$\Rightarrow$ dynamic programming (e.g. Value Iteration)

② If we know V^*, how to get π^* ?

$$\pi^*(s) = \underset{a \in A}{\arg\max} \left(R(s,a) + \gamma \cdot \mathbb{E}_{s' \sim p(\cdot|s,a)} \left[V^*(s') \right] \right)$$

$$\Downarrow$$

$$\boxed{\pi^*(s) = \underset{a \in A}{\arg\max}\, Q^*(s,a)}$$

$$Q^*(s,a) = R(s,a) + \gamma \cdot \mathbb{E}_{s' \sim p(\cdot|s,a)} \left[V^*(s') \right]$$

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

$\Rightarrow$ Bellman Optimality Equation (for Q^*):

② $$\boxed{Q^*(s,a) = R(s,a) + \gamma \cdot \mathbb{E}_{s' \sim p(\cdot|s,a)} \left[\max_{a' \in A} Q^*(s',a') \right]}$$

State-action pairs

$|S \times A|$ unknown variables

$|S \times A|$ equations

$$\Downarrow$$

$$\pi^*(s) = \underset{a \in A}{\arg\max}\, Q^*(s,a)$$

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \,\middle|\, s_1 = s \right]$$

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \,\middle|\, s_1 = s,\ a_1 = a \right]$$

Bellman Equation:

$$V^{\pi}(s) = R(s,\pi) + \gamma \cdot \mathbb{E}_{s' \sim P(\cdot|s,\pi)}\left[V^{\pi}(s') \right] \quad (3)$$

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

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

$$\Rightarrow Q^{\pi}(s,a) = R(s,a) + \gamma \cdot \mathbb{E}_{\substack{a \sim \pi(s) \\ s' \sim P(\cdot|s,a)}}\left[Q^{\pi}(s,a) \right] \quad (4)$$

↳ Q^{π} is the unique solution of BE.

State-action discounted occupancy $d^{\pi}(s,a)$

Now: prove the optimality theorem.

Thm Let Π be the set of all policies. < non-stationary (history dependent) / random.

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

There exists a stationary and deterministic policy $\pi \in \Pi$

s.t. $\forall s \in S, a \in A$: $V^\pi(s) = V^*(s)$, $Q^\pi(s,a) = Q^*(s,a)$

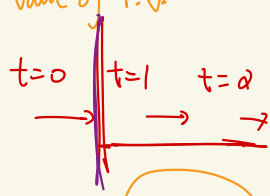
Proof: $\pi\big(A_t = a \mid \underbrace{S_0 = s_0, A_0 = a_0, \overset{R_0 = r_0}{\ldots}, S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \overset{R_{t-1} = r_{t-1}}{} S_t = s_t}_{\text{history}}\big)$

MDP ← Markovian

S, A: random variable

s, a: value of r.v.

$(S_0, A_0, R_0, S_1) = (s, a, r, s')$



$$\sup_{\pi \in \Pi} \mathbb{E}\Big[\sum_{t=1}^{\infty} \gamma^t r(S_t, a_t) \,\Big|\, \pi, \underbrace{(S_0, A_0, R_0, S_1) = (s, a, r, s')}_{\text{environment (Markov)}}\Big] = \gamma \cdot V^*(s')$$

π

time-shift argument

$\pi \to \pi_{(s,a,r)}$ "offset" policy

$\pi_{(s,a,r)}(A_t = a' \mid S_0 = s_0, A_0 = a_0, R_0 = r_0, \ldots, S_t = s_t)$

$$= \pi\left(A_{t+1}=a' \,\middle|\, S_0=s,\ A_0=a,\ R_0=r,\ S_1=s_0,\ A_1=a_0,\ R_0=r_0,\ \cdots;\ S_{t+1}=s_{t+1}\right)$$

Π : all non-stationary & random policies

$$\Pi_{(s,a,r)} = \{\pi_{(s,a,r)} : \pi \in \Pi\} = \Pi \qquad \forall (s,a,r)$$

$$\sup_{\pi\in\Pi} \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t r(S_t,A_t) \,\middle|\, \pi,\ (S_0,A_0,R_0,S_1)=(s,a,r,s')\right]$$

$$= \sup_{\pi\in\Pi} \gamma\cdot \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t\cdot r(S_t,A_t) \,\middle|\, \pi_{(s,a,r)},\ S_0=s'\right]$$

$$= \sup_{\pi\in\Pi} \gamma\cdot V^{\pi_{(s,a,r)}}(s') = \sup_{\pi\in\Pi_{(s,a,r)}} \gamma\cdot V^{\pi}(s')$$

$$= \gamma\cdot \sup_{\pi\in\Pi} V^{\pi}(s') = \gamma\cdot V^*(s')$$

deterministic & stationary policy $\tilde{\pi}$:

$$\boxed{\tilde{\pi}(s) = \sup_{a\in A} \mathbb{E}\left[r(s,a)+\gamma\cdot V^*(s') \,\middle|\, (S_0,A_0)=(s,a)\right]}$$

Optimality : $V^*(s_0) = \sup_{\pi\in\Pi} \mathbb{E}_{\pi}\left[r(s_0,a_0) + \sum_{t=1}^{\infty} \gamma^t\cdot r(s_1,a_1)\right]$

$$= \sup_{\pi\in\Pi} \mathbb{E}\left[r(s_0,a_0) + \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t r(s_1,a_1) \,\middle|\, \pi,\ (S_0,A_0,R_0,S_1)=(s_0,a_0,r_0,s_1)\right]\right]$$

$$\leq \sup_{\pi\in\Pi} \mathbb{E}\left[r(s_0,a_0) + \sup_{\pi'\in\Pi} \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t r(s_1,a_1) \,\middle|\, \pi',\ (S_0,A_0,R_0,S_1)=(s_0,a_0,r_0,s_1)\right]\right]$$

$$= \sup_{\pi \in \Pi} \mathbb{E}\left[r(s_0, a_0) + \gamma \cdot V^*(s_1) \right]$$

independent to π

$$= \sup_{a_0 \in A} \mathbb{E}\left[r(s_0, a_0) + \gamma \cdot V^*(s_1) \right]$$

$$= \mathbb{E}\left[r(s_0, a_0) + \gamma \cdot V^*(s_1) \,\middle|\, \tilde{\pi} \right]$$

$$V^*(s_0) \leq \mathbb{E}\left[r(s_0, a_0) + \gamma \cdot V^*(s_1) \,\middle|\, \tilde{\pi} \right]$$

$$\leq \mathbb{E}\left[r(s_0, a_0) + \gamma \cdot r(s_1, a_1) + \gamma^2 \cdot V^*(s_2) \,\middle|\, \tilde{\pi} \right]$$

$$\leq \cdots$$

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

On the other hand, $V^*(s_0) = \sup_{\pi \in \Pi} V^{\pi}(s_0) \geq V^{\tilde{\pi}}(s_0)$

$$\Rightarrow V^*(s_0) = V^{\tilde{\pi}}(s_0)$$

optimal policy $\tilde{\pi}$: satisfy "Bellman optimality equation for V^*. $\square$

greedy policy

Δ $V^*(s) = \max_{a \in A} Q^*(s, a)$

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

$$V^*(s) \geq Q^{\pi^*}(s, a) = Q^*(s, a) \quad \text{for any } a \in A$$