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

(s_1: r.v., s: realization)

S_t: r.v.
s, s': realizations

$$= \mathbb{E}\left[r_1 + \sum_{t=2}^{\infty} \gamma^{t-1} \cdot r_t \,\middle|\, s_1 = s, \pi\right]$$

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

$$= R(s, \pi(s)) + \gamma \mathbb{E}_{s_2}\left[\mathbb{E}\left[\sum_{t=2}^{\infty} \gamma^{t-2} \cdot r_t \,\middle|\, s_1 = s, s_2, \pi\right]\right]$$

(s_2: r.v.)

function of s_2.

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

$$\mathbb{E}\left[\sum_{t=2}^{\infty} \gamma^{t-2} \cdot r_t \,\middle|\, s_1 = s, s_2 = s', \pi\right].$$

$s_1, a_1, r_1, \boxed{s_2, a_2, r_2, s_3, a_3, r_3, \cdots}$

s $\quad$ s'

sum of discounted rewards

$$= \mathbb{E}\left[\sum_{t=2}^{\infty} \gamma^{t-2} \cdot r_t \,\middle|\, s_2 = s', \pi\right].$$

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

Bellman Eq. for Policy Evaluation

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

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

$$= R(s,\pi(s)) + \gamma \left\langle P(\cdot|s,\pi(s)),\, V^\pi(\cdot)\right\rangle.$$

$P(\cdot|s,\pi(s))$: $|S|\times 1$ vec. $V^\pi(\cdot)$: $|S|\times 1$ vec.

More general: for $\pi: S \to \Delta(A)$.

replace: $R(s,\pi(s)) \longrightarrow \mathbb{E}_{a\sim\pi(\cdot|s),\, r\sim R(s,a)}[r]$.

$s' \sim P(\cdot|s,\pi(s)) \longrightarrow a \sim \pi(\cdot|s),\ s' \sim P(\cdot|s,a)$.

In this course: will use notation $R(s,\pi)$. $s' \sim P(\cdot|s,\pi)$. for both deterministic & stoch policies.

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

Matrix form: define.

- V^π as $|S|\times 1$ vec $[V^\pi(s)]_{s\in S}$.
- R^π as $[R(s,\pi)]_{s\in S}$.
- P^π as $|S|\times|S|$ matrix $[P(s'|s,\pi)]_{s\in S,\, s'\in S}$.

P^π: transition matrix of MC induced by π.

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

unknown. $\quad$ unknown.

$[P(\cdot|s,\pi)]^T$ ← $\begin{bmatrix} \text{---} \end{bmatrix} \begin{bmatrix} | \\ V^\pi \\ | \end{bmatrix} = \begin{bmatrix} \cdot \end{bmatrix}_S$.

Solve: $V^\pi - \gamma P^\pi V^\pi = R^\pi$.

$(I - \gamma P^\pi) V^\pi = R^\pi$. → does inv. always exist?

$V^\pi = \underbrace{(I - \gamma P^\pi)^{-1}}_{\Delta} R^\pi$.

YES! To prove. w.t.s. $I - \gamma P^\pi$ is non-singular.

matrix A invertible $\iff \forall x \neq \vec{0}. \; Ax \neq \vec{0}$.

Consider arbitrary $x \neq \vec{0}$. w.t.s. $(I - \gamma P^\pi) x \neq \vec{0}$.

$x \neq \vec{0} \iff \|x\|_\infty > 0$.

$\|(I - \gamma P^\pi) x\|_\infty = \|x - \gamma P^\pi x\|_\infty$.

$\geq \|x\|_\infty - \|\gamma P^\pi x\|_\infty$ (triangular ineq. for norms).

$= \|x\|_\infty - \gamma \cdot \|P^\pi x\|_\infty$.

$= \|x\|_\infty - \gamma \cdot \max_s |\langle P(\cdot|s,\pi), x \rangle|$.

$\geq \|x\|_\infty - \gamma \cdot \|x\|_\infty$

$= (1-\gamma) \cdot \|x\|_\infty > 0$.

Def. $\|\cdot\|_\infty$.

$\|f\|_\infty = \max_s |f(s)|$.

apply Hölder's ineq.

$|\langle u, v \rangle| \leq \|u\|_1 \cdot \|v\|_\infty$

$u = P(\cdot|s,\pi)$

$\|u\|_1 = 1$.

$v = x$.

$\|v\|_\infty = \|x\|_\infty$.

$\begin{bmatrix} P(\cdot|s,\pi) \\ \Delta \end{bmatrix} \begin{bmatrix} | \\ x \\ | \end{bmatrix}$.

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

$\begin{cases} R^\pi = [R(s,\pi)]_{s\in S} \\ P^\pi = [P(s'|s,\pi)]_{s,s'\in S} \end{cases}$

▷ $(1-\gamma)\cdot(I-\gamma P^\pi)^{-1} \in \mathbb{R}^{|S|\times|S|}$.

the row vector indexed by s. is

normalized discounted occupancy of π, $(d^{\pi,s})^T \in \mathbb{R}^{|S|}$

$$d^{\pi,s}(s') = (1-\gamma)\mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} \cdot \mathbb{I}[\underbrace{s_t = s'}_{\Delta}] \,\middle|\, s_1 = s, \pi\right].$$

Equivalent def: $d_t^{\pi,s}(s') = \mathbb{E}[\mathbb{I}[s_t = s'] \,|\, s_1 = s, \pi]$.

marginal dist. of s_t, under π, starting from s.

$= \mathbb{P}[s_t = s' \,|\, s_1 = s, \pi]$.

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

$$(I-\gamma P^\pi)^{-1} = [(d^{\pi,s})^T]_{s\in S}.$$

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

$$V^\pi(s) = \underbrace{e_s^T \cdot (I-\gamma P^\pi)^{-1}}_{\uparrow} R^\pi = \frac{1}{1-\gamma}\cdot (\underbrace{d^{\pi,s}}_{\Delta})^T R^\pi$$

$[0,0,\cdots,0,\underbrace{1}_{s\text{-th element}},0,\cdots,0]^T$

$$= \frac{1}{1-\gamma}\cdot \mathbb{E}_{\tilde{s}\sim d^{\pi,s}}[R^\pi(\tilde{s})]$$

$$= \frac{1}{1-\gamma}\cdot \mathbb{E}_{\tilde{s}\sim d^{\pi,s}}[R(\tilde{s},\pi)].$$

Policy Optimization

Thm. (see Puterman '94). For infinite-horizon discounted MDPs.
there always exists a stationary & deterministic policy
that is optimal for all starting states simultaneously.
denote such a policy as π^*.

Def. $V^* := V^{\pi^*}$.

Bellman Optimality Eq. $\forall s \in S$.

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

what if we take a in s. $Q^*(s,a)$.

argmax = optimal action!!

$$V^*(s) = \max_a Q^*(s,a).$$

Remarks:

1. $V^\pi = R^\pi + \gamma P^\pi V^\pi$ is linear. but optimality eq. is non-linear.
2. Solution exists & is unique (will show later).
   but π^* is not unique

Q: how to get π^* given V^*??

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

if there are ties, π^* is not unique.
can choose any action in the tie,
or even an distribution over optimal actions.

For convenience: Optimal Q-function (or action-value func.)

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

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

Bellman Optimality Eq. (for Q^*)

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

Remarks:

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

(π: unrestricted)

$$Q^*(s,a) = \max_{\pi} \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid s_1 = s, a_1 = a, a_{2:\infty} \sim \pi\right]$$

2. Bellman opt. eq: $V^*(s) = \max_{a \in A}\left(R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]\right)$

Bellman eq for PE: $V^\pi(s) = [R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^\pi(s')]]$ evaluate at $a = \pi(s)$.

3. can also define Q^π (analogues to Q^*)

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

* compare to Bellman eq. for Q^*

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

$$Q^\pi(s,a) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid s_1 = s, a_1 = a, a_{2:\infty} \sim \pi\right]$$

How to solve for V^* / Q^* | Value Iteration (V2)

VI for solving Q^*:

func ↓ func ↓

- Definition: Bellman optimality operator $\mathcal{T}: \mathbb{R}^{S\times A} \to \mathbb{R}^{S\times A}$.

$\forall f \in \mathbb{R}^{S\times A}$, $(\mathcal{T}f) \in \mathbb{R}^{S\times A}$.

$$(\mathcal{T}f)(s,a) = R(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\max_{a'} f(s',a')\right]$$

Bellman opt. eq. for Q^*: $Q^* = \mathcal{T}Q^*$

(Q^* is the fixed point of operator $\mathcal{T}$).