Lecture - 03

- Reproducibility
- Ablation / Analysis.
- Applications.

} 3 per team

- Research

} 1 - 3 per team

Finite Markov Decision process

S A R

S → States
A → Actions
R → Rewards

$$p(s', r \mid s, a)$$ dynamics fn.

Markov Property

$$p(S_t, R_t \mid S_{t-1}, A_{t-1})$$

Return

episodic problems.

$$G_t = R_{t+1} + R_{t+2} + \cdots + R_T$$

Continuing tasks

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

$$= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

$0 \leq \gamma \leq 1$ is called the discount rate

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

$$= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots$$

$$= R_{t+1} + \gamma \left(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} \cdots \right)$$

$$G_t = R_{t+1} + \gamma G_{t+1}$$

Note: In Continuing tasks,

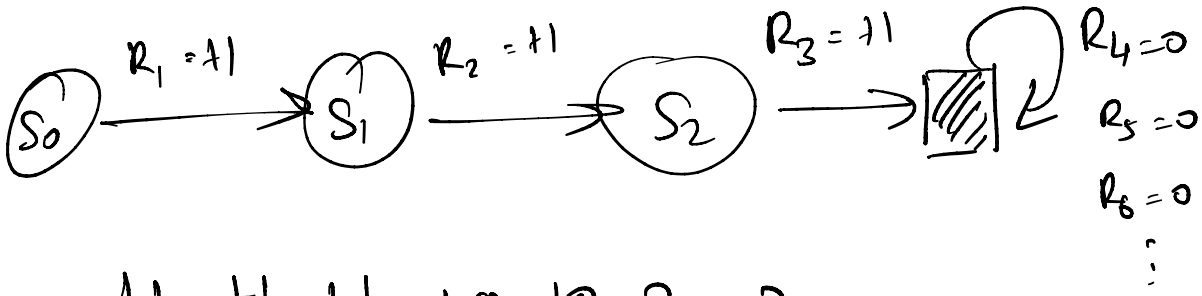
Sum of ~~rewards is~~ infinite number of rewards, the sum is still finite.

ex: Constant +1 reward.

$$G_t = \sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}$$

Episodic | Continuing.

$G_t = R_{t+1} + R_{t+2} + \dots + R_T$ $\quad$ $G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$



$+1, +1, +1, +0, +0, 0, 0 \dots$

$$G_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k .$$

$T = \infty$ or $\gamma = 1$

MDP, dynamics, returns.

Policies and Value fns:-

Goal: estimate value fns.

Value fn
- State
- State-action pairs.

Policy: mapping from states to prob. of selecting each possible action.

$$\pi(a|s)$$

State value fn:

$V_\pi(s)$ = value fn of a state S under a policy π.

$$V_\pi(s) = \mathbb{E}_\pi \left[G_t \mid S_t = s \right]$$

$$= \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s \right] \quad \forall s \in S$$

Note: V_π (terminal state) $= 0$.

Action-value fn:

$q_\pi(s,a)$ = value of taking action a in state s.

$$v_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s]$$

$$q_\pi(s,a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$$

$$= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a\right]$$

Bellman equation for value fns:-

$$V_\pi(s) = \mathbb{E}_\pi\left[G_t \mid S_t = s\right]$$

$$= \mathbb{E}_\pi\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s\right]$$

$$= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r | s, a)$$

$$\left[r + \gamma \mathbb{E}_\pi\left[G_{t+1} \mid S_{t+1} = s'\right] \right]$$

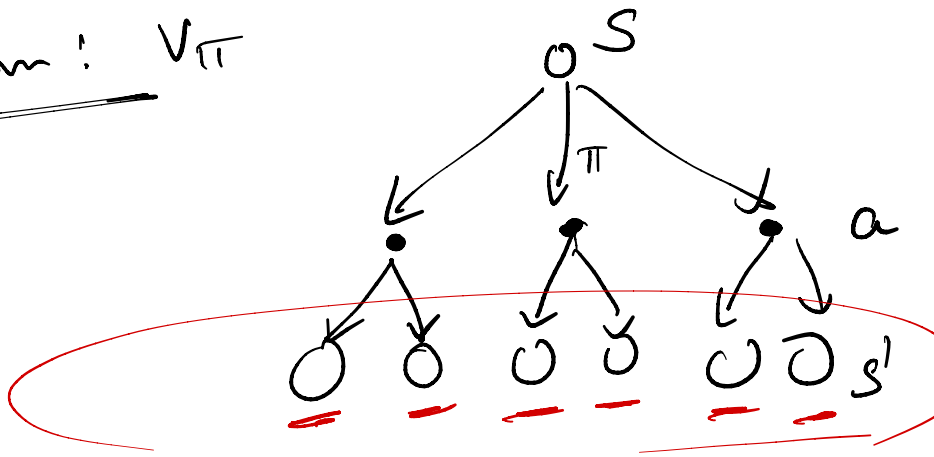
Bellman equation:

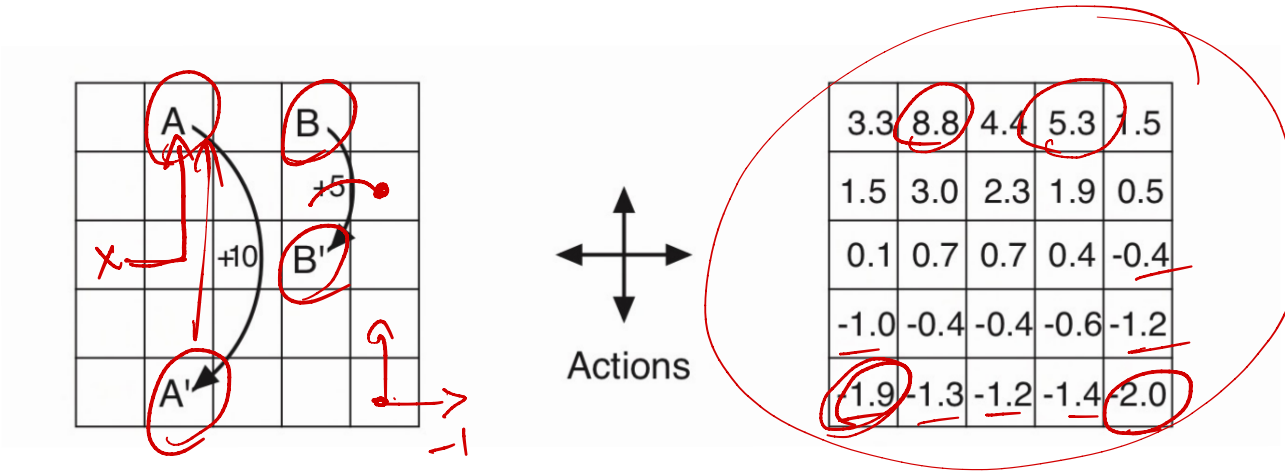
$$V_\pi(s) = \sum_a \pi(a|s) \sum_{s', r} p(s', r | s, a)\left[r + \gamma V_\pi(s')\right]$$

$\forall s \in S$

Unique soln: V_π.

backup diagram: V_π

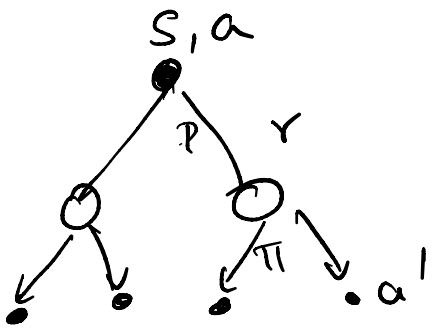




equiprobable random policy

Bellman eqn for $q_\pi(s,a)$:-

$$q_\pi(s,a) = \sum_{s',r} p(s',r|s,a)\left[r + \gamma \sum_{a'} \pi(a'|s') q_\pi(s',a')\right]$$



Optimal policies / optimal value fn.

$\pi \geq \pi'$ if and only if $V_\pi(s) \geq V_{\pi'}(s)$ $\forall s \in S$.

Optimal policy: π_*.

'n' states
'k' actions.
n^k.

V_* — optimal state value fn.

$$V_*(s) = \max_\pi V_\pi(s) \quad \forall s \in S$$

$$q_*(s,a) = \max_\pi q_\pi(s,a) \quad \forall s \in S, \; a \in A$$

$$q_*(s,a) = \mathbb{E}\left[R_{t+1} + \gamma V_*(S_{t+1}) \,\middle|\, S_t = s, A_t = a \right]$$

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

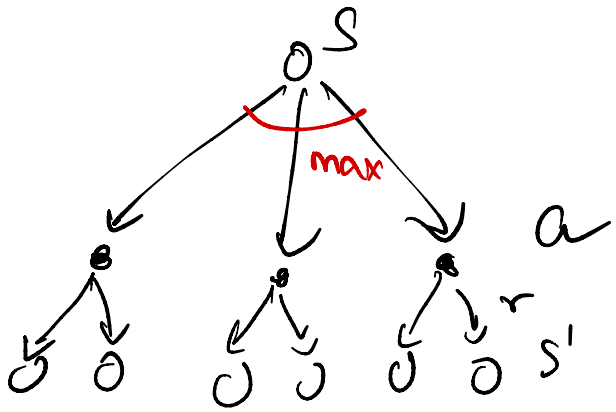
$$= \max_a \mathbb{E}_{\pi_*}\left[G_t \mid S_t = s, A_t = a \right]$$

Bellman optimality eqn for v_*:

$$= \max_a \mathbb{E}_{\pi_*}\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a \right]$$

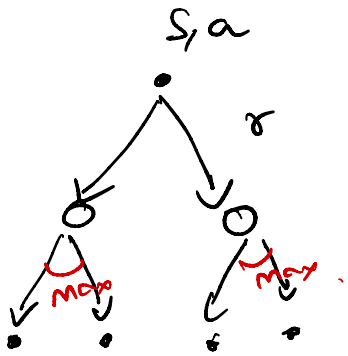
$$V_*(s) = \max_a \mathbb{E}\left[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a \right]$$

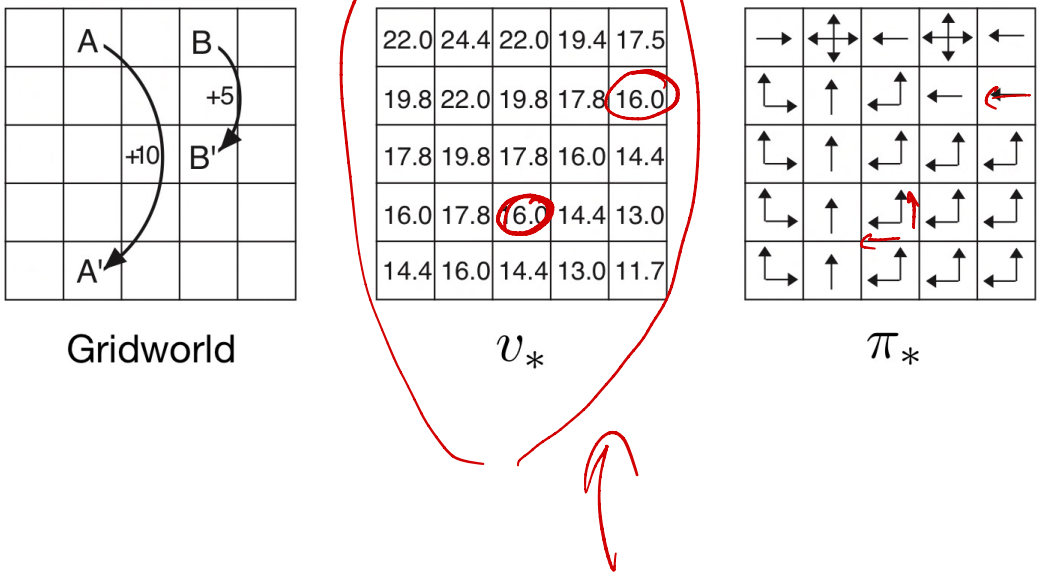
$$V_*(s) = \max_a \sum_{s', r} p(s', r \mid s, a)\left[r + \gamma v_*(s') \right]$$



$$q_*(s,a) = \mathbb{E}\left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \,\middle|\, S_t = s, A_t = a \right]$$

$$= \sum_{s',r} P(s', r \mid s, a) \left[r + \gamma \max_{a'} q_*(s', a') \right]$$





Gridworld v_* π_*

Strong assumption !

(1) We accurately know the dynamics.

(2) We have enough Compute to complete the Sish.

(3) Markov property.

Dynamic Programming

Given an MDP, compute the optimal policy.

Finite MDP $p(s', r | s, a)$.

Key idea: Use the value fns to organize and structure the search for good policies

Bellman Optimality equations:

$$v_*(s) = \max_a \mathbb{E}\left[R_{t+1} + \gamma v_*(S_{t+1}) \,\middle|\, S_t = s, A_t = a \right]$$

$$= \max_a \sum_{s', r'} p(s', r | s, a) \left[r + \gamma v_*(s') \right]$$

$$q_*(s, a) = \mathbb{E}\left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \,\middle|\, S_t = s, A_t = a \right]$$

$$= \sum_{s' r} p(s', r | s, a) \left[r + \gamma \max_{a'} q_*(s', a') \right]$$

given π, how to estimate $V_\pi(s)$?

given $V_\pi(s)$, how to find a better policy π' ?

Policy evaluation (prediction)

Policy improvement

Policy evaluation:

Given a arbitrary policy π, how to compute V_π ?

$$V_\pi(s) = \mathbb{E}[G_t \mid S_t = s]$$

$$= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s]$$

$$= \mathbb{E}_\pi[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s]$$

$$V_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\left[r + \gamma V_\pi(s')\right]$$

(*)

Iterative soln:

Expected update

$V_0, V_1, V_2, V_3 \cdots$

$$V_{k+1}(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\left[r + \gamma V_k(s')\right]$$

$V_k(s) = V_{k+1}(s)$ V_0 (_____)

vs. $V_k = V_\pi$ V_1 ( )

"Sweep" ⟷ iteration.

$V_k \longrightarrow V_\pi$ when $k \to \infty$

if $V_{k+1}(s) = V_k(s) \quad \forall s.$

$$V_k = V_{k+1} = V_\pi.$$

V_π - is the fixed pt for this update rule.

$$\{V_k\} \longrightarrow V_\pi$$

as $k \to \infty$.

Iterative Policy evaluation

$V_k(s)$:

$V_{k+1}(s)$:

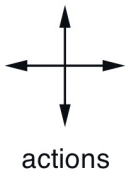
Iterative Policy Evaluation, for estimating $V \approx v_\pi$

Input π, the policy to be evaluated
Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

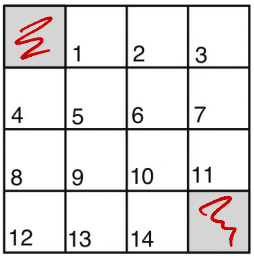
Loop:
  $\Delta \leftarrow 0$
  Loop for each $s \in \mathcal{S}$:
    $v \leftarrow V(s)$
    $V(s) \leftarrow \sum_a \pi(a|s) \sum_{s',r} p(s', r \,|\, s, a)\big[r + \gamma V(s')\big]$
    $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

(Sweep)

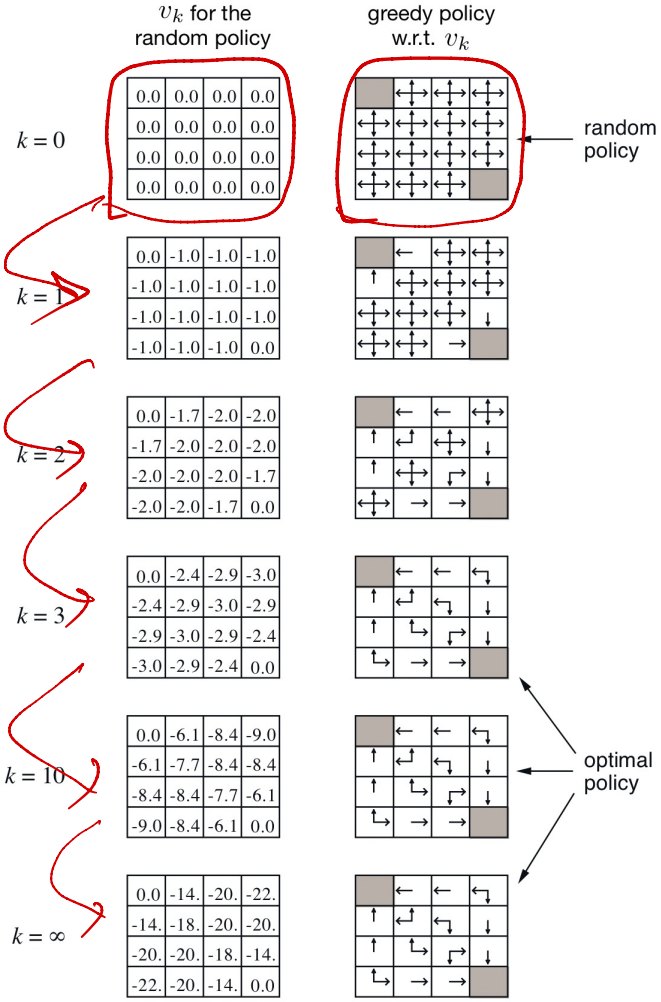
$\theta = 0.1$



actions



$R_t = -1$
on all transitions



Policy improvement :-

Compute Val. fns for a policy to help find better policies

π, V_π

's' $a = \pi(s)$ or $a \neq \pi(s)$.

Consider selecting 'a' in 's' and thereafter following the existing pol^y π.

$$q_\pi(s,a) = \mathbb{E}\left[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s, A_t = a \right]$$

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

if $q_\pi(s,a) > V_\pi(s)$

Policy improvement theorem :- (PIT)

Let π and π' be a pair of deterministic policies such that

$\forall\ s \in S$

$$q_\pi(s, \pi'(s)) \geq v_\pi(s)$$

then the policy π' must be as good as or better than π.

i.e. $v_{\pi'}(s) \geq v_\pi(s)\ \forall s \in S$

Proof:

$$v_\pi(s) \le q_\pi(s, \pi'(s))$$

$$= \mathbb{E}\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t = \pi'(s) \right]$$

$$= \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s \right]$$

$$\le \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma q_\pi(S_{t+1}, \pi'(S_{t+1})) \mid S_t = s \right]$$

$$= \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma \mathbb{E}_{\pi'}\left[R_{t+2} + \gamma v_\pi(S_{t+2}) \mid S_{t+1}, A_{t+1} = \pi(S_{t+1}) \right] \mid S_t = s \right]$$

$$= \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 v_\pi(S_{t+1}) \mid S_t = s \right]$$

$$\le \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 v_\pi(S_{t+3}) \mid S_t = s \right]$$

$$\vdots$$

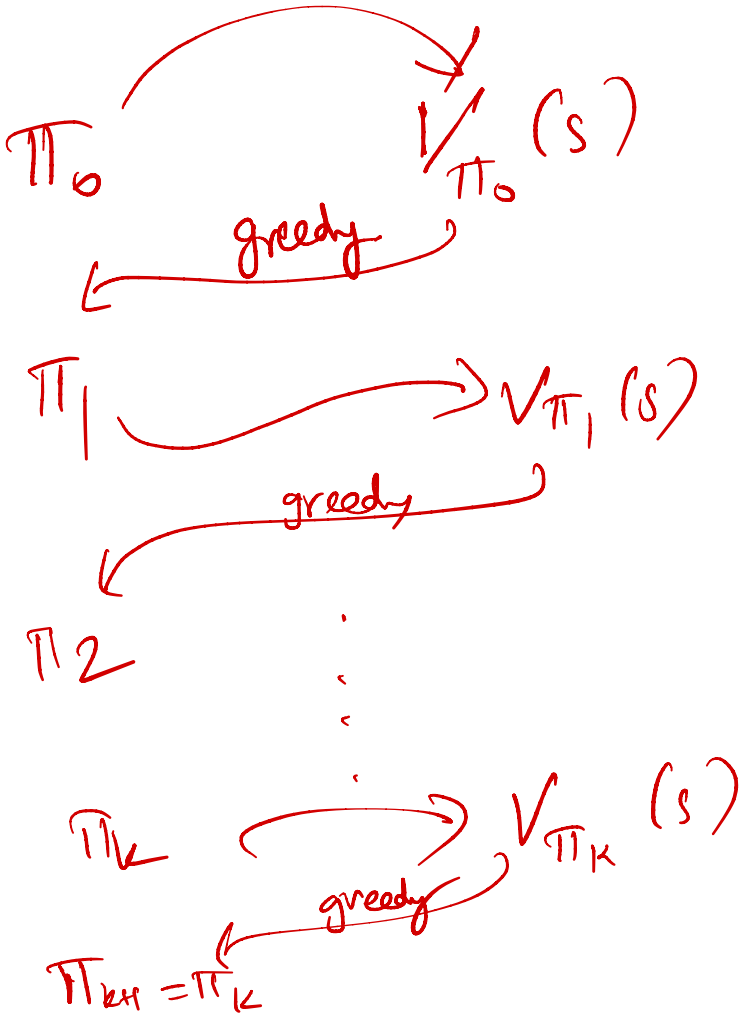
$$\le \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \dots \right]$$

$$= v_{\pi'}(s)$$

Given a value fn. for some policy π, new policy π' wrt. val. fn.

Greedy policy.

$$\pi'(s) = \arg\max_a q_\pi(s,a)$$



$$V_\pi = V_{\pi'}$$

$$V_{\pi'}(s) = \max_a \mathbb{E}\left[R_{t+1} + \gamma V_{\pi'}(S_{t+1}) \mid S_t = s, A_t = a \right]$$

$$V_{\pi'} = V_*$$

$$\pi_0 \xrightarrow{E} V_{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} V_{\pi_1} \xrightarrow{I} \pi_2 \cdots \pi_* \xrightarrow{E} V_*$$

Policy iteration!

Policy Iteration (using iterative policy evaluation) for estimating $\pi \approx \pi_*$

1. Initialization
   $V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$

2. Policy Evaluation
   Loop:
      $\Delta \leftarrow 0$
      Loop for each $s \in \mathcal{S}$:
         $v \leftarrow V(s)$
         $V(s) \leftarrow \sum_{s',r} p(s', r|s, \pi(s))\big[r + \gamma V(s')\big]$
         $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
   until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

3. Policy Improvement
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
      old-action $\leftarrow \pi(s)$
      $\pi(s) \leftarrow \arg\max_a \sum_{s',r} p(s', r|s, a)\big[r + \gamma V(s')\big]$
      If *old-action* $\neq \pi(s)$, then *policy-stable* $\leftarrow$ *false*
   If *policy-stable*, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2

Value iteration:

$$v_{k+1}(s) = \max_a \mathbb{E}\big[R_{t+1} + \gamma v_k(S_{t+1}) \,\big|\, S_t = s, A_t = a\big]$$

$$= \max_a \sum_{s',r} p(s', r|s, a)\big[r + \gamma v_k(s')\big] \quad \forall s \in \mathcal{S}$$

Value Iteration, for estimating $\pi \approx \pi_*$

Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(\mathit{terminal}) = 0$

Loop:
| $\Delta \leftarrow 0$
| Loop for each $s \in \mathcal{S}$:
| $\quad v \leftarrow V(s)$
| $\quad V(s) \leftarrow \max_a \sum_{s',r} p(s', r \,|\, s, a)\big[r + \gamma V(s')\big]$
| $\quad \Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

Output a deterministic policy, $\pi \approx \pi_*$, such that
$\pi(s) = \operatorname{argmax}_a \sum_{s',r} p(s', r \,|\, s, a)\big[r + \gamma V(s')\big]$

First-visit MC prediction, for estimating $V \approx v_\pi$

Input: a policy π to be evaluated

Initialize:

$V(s) \in \mathbb{R}$, arbitrarily, for all $s \in \mathcal{S}$

$Returns(s) \leftarrow$ an empty list, for all $s \in \mathcal{S}$

Loop forever (for each episode):

Generate an episode following π: $S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow \gamma G + R_{t+1}$

Unless S_t appears in $S_0, S_1, \ldots, S_{t-1}$:

Append G to $Returns(S_t)$

$V(S_t) \leftarrow \text{average}(Returns(S_t))$

Monte Carlo ES (Exploring Starts), for estimating $\pi \approx \pi_*$

Initialize:

$\pi(s) \in \mathcal{A}(s)$ (arbitrarily), for all $s \in \mathcal{S}$

$Q(s, a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

$Returns(s, a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Loop forever (for each episode):

Choose $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ randomly such that all pairs have probability > 0

Generate an episode from S_0, A_0, following π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow \gamma G + R_{t+1}$

Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:

Append G to $Returns(S_t, A_t)$

$Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$

$\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$

On-policy first-visit MC control (for ε-soft policies), estimates $\pi \approx \pi_*$

Algorithm parameter: small $\varepsilon > 0$

Initialize:

$\pi \leftarrow$ an arbitrary ε-soft policy

$Q(s, a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

$Returns(s, a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Repeat forever (for each episode):

Generate an episode following π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow \gamma G + R_{t+1}$

Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:

Append G to $Returns(S_t, A_t)$

$Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$

$A^* \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken arbitrarily)

For all $a \in \mathcal{A}(S_t)$:

$$\pi(a|S_t) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(S_t)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(S_t)| & \text{if } a \neq A^* \end{cases}$$