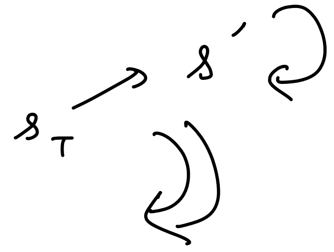
Reinforcement Learning.

- States s_t — transition function $p(s_{t+1}, r_{t+1} \mid s_t, a_t)$
- Actions a_t
- Episodes $s_0\, a_0,\ s_1\, a_1\, \ldots\ldots$
  - Episodic tasks $s_0 \ldots\ldots$
  - Continuing tasks $s_0 \ldots\ldots$



Returns $G_t = R_{t+1} + R_{t+2} + \ldots\ldots$

Infinite continuing episodes, introduce a discount factor γ

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

$$0 \le \gamma < 1$$

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

- Policy $\pi(a \mid s)$ = Probability of executing action a when in state s.

- Value function under policy π, $v_\pi(s) \rightarrow \mathbb{R}$

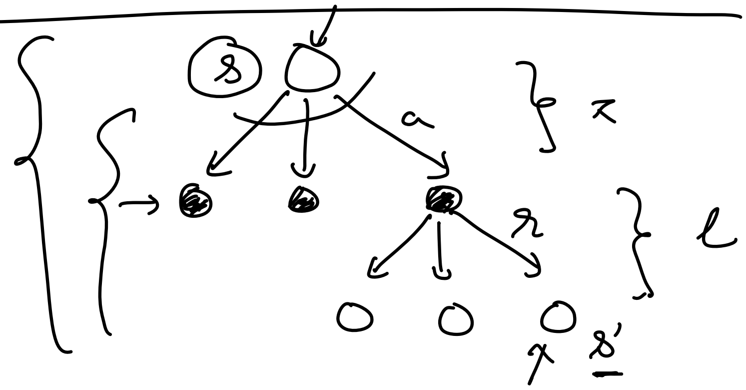
  expected return when following policy π starting from state s.

$$v_\pi(s) = E_\pi\left[G_t \mid S_t = s\right]$$

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

$$= \sum_a \pi(a \mid s) \sum_{s', r} p(s', r \mid s, a)\left[r + \gamma E_\pi\left[G_{t+1} \mid S_{t+1} = s'\right]\right]$$

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



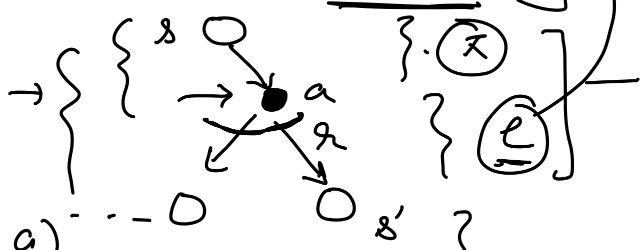
Action Value Function

$q_\pi(s,a) \doteq$ Expected return if I were to execute action a from state s & then execute π.

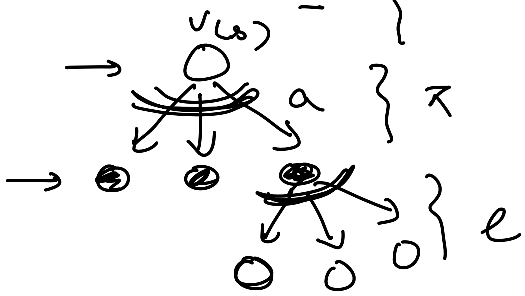
$$q_\pi(s,a) = E_\pi\left[G_t \mid S_t = s, A_t = a\right]$$

$$= E_\pi\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a\right]$$

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



$$v_\pi(s) = \sum_a \pi(a|s)\, q_\pi(s,a)$$



Optimal value function

$$v_*(s) \doteq \max_\pi v_\pi(s)$$

$$q_*(s,a) \doteq \max_\pi q_\pi(s,a)$$

<u>Optimal policies.</u>

$\pi \geq \pi'$ if & only if $v_\pi(s) \geq v_{\pi'}(s) \quad \forall s \in S$

π^* is optimal if $\pi^* \geq \pi \quad \forall \pi$

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

<u>Bellman Optimality Equation</u>

$$\begin{aligned} v_*(s) &= \max_a q_*(s,a) \\ &= \max_a E_{\pi^*}\left[G_t \mid S_t = s, A_t = a\right] \\ &= \max_a E_{\pi^*}\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a\right] \\ &= \max_a E\left[R_{t+1} + \gamma v_{\pi^*}(s_{t+1})\right] \end{aligned}$$

$$\boxed{v_*(s) = \max_a \sum_{s',r} p(s', r \mid s,a)\left[r + \gamma v_*(s')\right]} \quad (1)$$

- if I give you $v_*(s)$, can you act in this MDP?

If instead, I gave you $q^*(s,a)$, can you act?

$$\leftarrow \arg\max_a q^*(s,a)$$

<u>Solving Known MDPs.</u>

- Policy evaluation

given a policy π, compute v_π

- Policy Improvement

given a policy π, obtain an improved policy π'

- Policy Iteration
- Value Iteration

Policy Evaluation

$|S|$

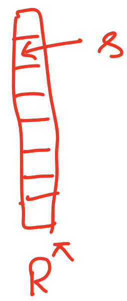
Given π. I want to evaluate $v_\pi(s)$

(a) Solving a system of equations

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

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

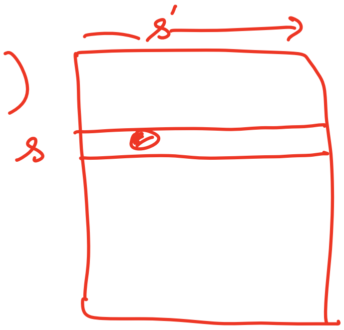
$R^\pi(s)$



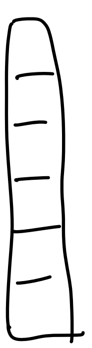
$$+ \gamma \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\, v_\pi(s')$$

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

$T^\pi(s, s')$



$$V_\pi = R^\pi + \gamma T^\pi V_\pi$$



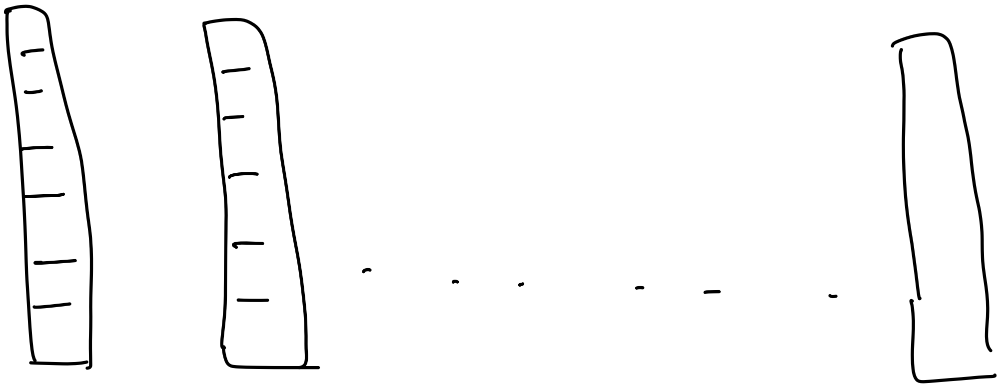
$$V_\pi = (I - \gamma T^\pi)^{-1} R^\pi .$$

(b) Policy Iteration.

- Initialize $v_0(s) = 0$
- for $k = 1 \ldots N$

  for $s \in S$

  $$\underline{v_{k+1}(s)} \leftarrow \underline{\sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\left[r + \gamma v_k(s')\right]}$$



$V_0(s)$ $\quad$ $V_1(s)$ $\quad \ldots \quad$ $V_\pi(s)$

γ- contraction

$$\Rightarrow \|V_{k+1} - U_{k+1}\|_\infty \le \gamma \, \|V_k - U_k\|_\infty$$

$$V_{k+1} \leftarrow R^\pi + \gamma T^\pi V_k$$

Policy Improvement

$\pi, \; v_\pi(s) \longrightarrow \pi'$

Policy Improvement Theorem

Given a value function for a policy π, we can obtain a greedifted policy π':

$$\rightarrow \underline{\pi'(s)} = \underset{a}{\operatorname{argmax}} \sum_{s',r} p(s',r|s,a)\left[r + \gamma v_\pi(s')\right]$$

π' better than π

$$\underline{v_{\pi'}(s)} \geq v_\pi(s) \quad \forall s$$

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

$$\underline{v_\pi(s)} \leq \underline{q_{\underline{\pi}}(s, \pi'(s))} \quad \cdots - . \quad v_{\pi'}(s)$$

$$= E_{\underline{\pi}}[\underline{\underline{R_{t+1}} + \gamma \underline{G_{t+1}}} \mid S_t = s, \underline{A_t = \pi'(s)}] \quad E_{\pi'}$$

$$= E_{\pi'}[R_{t+1} + \gamma v_{\pi}(S_{t+1}) \mid S_t = s]$$

$$\leq E_{\pi'}[R_{t+1} + \gamma \underline{q_\pi(S_{t+1}, \pi'(S_{t+1}))} \mid S_t = s]$$

$$= E_{\pi'}[R_{t+1} + \gamma E_\pi[R_{t+2} + \gamma v_\pi(S_{t+2}) \mid S_{t+1}, \underline{A_{t+1}} = \pi'(S_{t+1})] \mid S_t = s]$$

$$= E_{\pi'}[R_{t+1} + \gamma E_{\underline{\pi'}}[R_{t+2} + \gamma v_\pi(S_{t+2}) \mid \cdots]$$

$$\leq E_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 \cdots]$$

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

for $\pi'(s) \leftarrow \underset{a}{\operatorname{argmax}}\ q_\pi(s, a)$

$$v_{\pi'}(s) \geq v_\pi(s) \quad \forall s$$

π' is an improvement over π.

Suppose new policy π' is only as good as

π' then

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

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

Policy Iteration

$$\pi_0 \xrightarrow{E} v_{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} v_{\pi_1} \xrightarrow{I} \cdots\cdots \rightarrow v_*$$

Value Iteration

Turns out, one can be fairly lazy & only partialy complete polity iteration steps, & things still work out.

Loop

$\Delta \leftarrow 0$

for each $s \in S$

$v \leftarrow V(s)$

$$V(s) \leftarrow \max_a \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma V(s')\right]$$

$\Delta \leftarrow \max(\Delta, |V - v(s)|)$

until $\Delta < \epsilon$

Model-free policy evaluation
Model-free control

Model-free policy evaluation

Estimate $V_\pi(s)$ using a monte carlo estimate

$$E[G_t \mid S_t = s]$$

First Visit Monte Carlo

Every Visit Monte Carlo

Sample episodes.

within an episode,

the first time you visit a state s

$$\begin{cases} N(s) \leftarrow N(s)+1 \\ S(s) \leftarrow S(s) + G_t \end{cases}$$

G_t is return from this point onwards from the first time you visited S_t.

$$V(s) \leftarrow S(s)/N(s)$$

Incremental Update

$$\begin{cases} N(s_t) \leftarrow N(s_t)+1 \\ V(S_t) \leftarrow V(s_t) + \frac{1}{N(s_t)}\left(G_t - V(S_t)\right) \end{cases}$$

$$V(S_t) \leftarrow V(S_t) + \alpha\left(G_t - V(S_t)\right) -$$

Nudge my $V(S_t)$ towards G_t.

Temporal Difference Learning TD TD(0)

$$V(s_t) \leftarrow V(s_t) + \alpha\left(\underline{R_{t+1} + \gamma G_{t+1}} - V(S_t)\right)$$

$$\longrightarrow \quad V(s_t) \leftarrow V(s_t) + \alpha\left(\underbrace{\underline{R_{t+1} + \gamma V(S_{t+1})} - V(S_t)}_{\text{TD error.}}\right)$$

$(s_t, a_t, s_{t+1}, r_{t+1})$ TD target.

an estimate of my return based on current value function.

Model free policy evaluation

<u>TD</u>

- Can start learning from incomplete episodes.
- low variance updates but they can be <u>biased</u>.

$$V(s_t) \leftarrow V(s_t) + \alpha\left(R_{t+1} + \gamma V(S_{t+1}) - V(s_t)\right)$$

- u

Monte Carlo

high variance estimates.

$$\sum_{i=} R_{t+i}$$

but it is an unbiased estimate.