Model: predict what env will do next

Value func: $V_\pi(s) = E_\pi[R_{t+1} + \gamma R_{t+2} \cdots | S_t = s]$
prediction of future reward

State trans prob(P) = $P(s'|s,a)$
• state trans matrix: [# of states X # of states]

Reward(R): $R_s = E[R_{t+1} | S_t = s]$

Return(G) = $\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$

Markov reward process $<S, P, R, \gamma>$

Bellman eq: = immediate reward + successor reward

$V(s) = E[R_{t+1} + \gamma V(S_{t+1}) | S_t]$

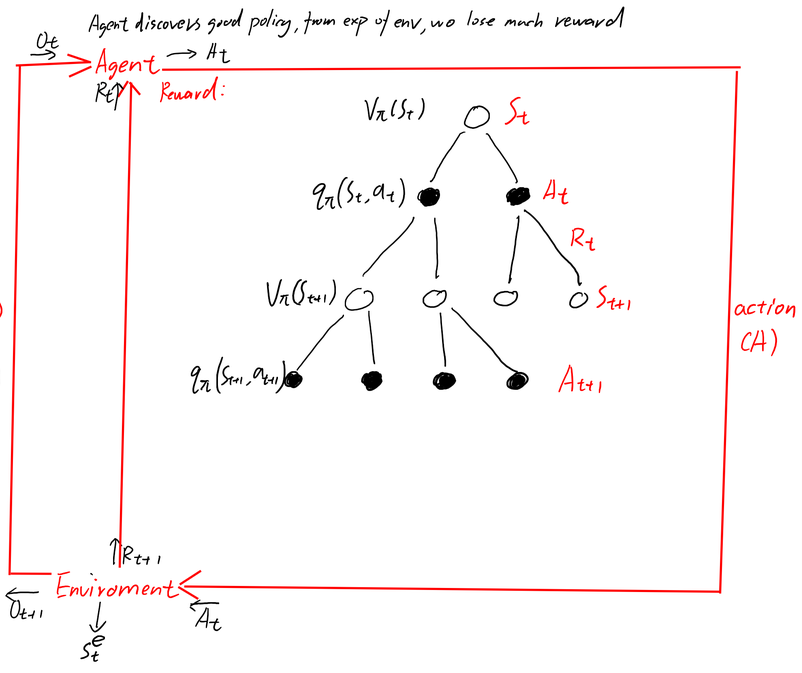
$= R + \gamma P V_{t+1} = (I - \gamma P)^{-1} R$

$O(\#\text{ of state's}^3)$ DP

=> finite states => MC, TD

infinite states => policy gradient

Agent discovers good policy, from exp of env, w/o lose much reward



Policy(π): map from state to action; deterministic: $a = \pi(s)$; stochastic: $\pi(a|s) = P(A_t | S_t)$
strict sense stationary (time independent)

Markov decision process $<S, P, R, \gamma, A>$

Only depends on current state

$P^\pi_{ss'} = \sum_{a \in A} \pi(a|s) P^a_{ss'}$, $R^\pi_s = \sum_{a \in A} \pi(a|s) R^a_s$

State value func:

expected return following policy π

$V_\pi(s) = E_\pi[G_t | S_t] = E_\pi[R_{t+1} + \gamma V_\pi(S_{t+1}) | S_t]$

$= \sum_{a \in A} \pi(a|s) \left(R^a_s + \gamma \sum_{s' \in S} P^a_{ss'} V_\pi(s') \right)$

action value func:

expected return following action then policy

$q_\pi(s,a) = E_\pi[G_t | S_t, A_t] = E_\pi\left[R_{t+1} + \gamma q(S_{t+1}, a_{t+1}) \middle| \begin{matrix} S_t \\ A_t \end{matrix}\right]$

$= R^a_s + \gamma \sum_{s' \in S} P^a_{ss'} \sum_{a' \in A} \pi(a'|s') q_\pi(s', a')$