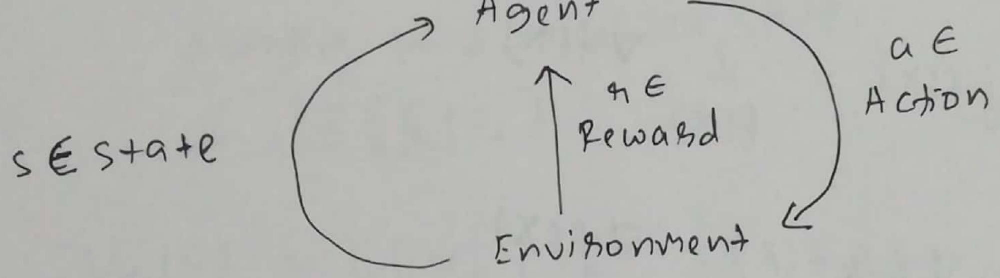
Markov Decision Process



$P(S_{t+1} \mid S_t, a_t)$: Formalization

$W_0 \leftarrow$ random

loop:

sample N sessions

elite sessions

$$W_{i+1} = W_i + \alpha \nabla \Big[\sum_{s_i, a_i \in Elite} \log \pi W_i (a_i \mid s_i) \Big]$$

RNN:

$$s_t^a = \sigma(s_{t-1}^a W_s + o_t W_o)$$

Expected Reward

$$J = \int N(\theta \mid \mu, \sigma^2) \int P(\tau \mid \theta) R(\tau) \, d\tau \, d\theta$$

θ = state τ = Action

Maximizing $J = \nabla J = \frac{\partial J}{\partial ?}$

So, $\boxed{\nabla J = \frac{\partial J}{\partial \mu \partial \sigma^2}}$

$$\nabla J = \int \nabla[N(\theta|\mu,\sigma^2)] \int P(\tau|\theta) R(\tau) \, d\tau \, d\theta$$

$$\nabla \log f(x) = \frac{1}{f(x)} \nabla f(x)$$

$$f(x) \nabla \log f(x) = \nabla f(x)$$

$$\nabla N(\theta|\mu,\sigma^2) = N(\theta|\mu,\sigma^2) \nabla \log(N(\theta|\mu,\sigma^2))$$

$$\nabla J = \int N(\theta|\mu,\sigma^2) \nabla \log(N(\theta|\mu,\sigma^2)) \int P(\tau|\theta) R(\tau) \, d\tau \, d\theta$$

$$\nabla J \approx \frac{1}{N} \sum_{i=0}^{N} \nabla \log N(\theta|\mu,\sigma^2) \sum R(s,a..)$$

strategy for optimization:

① Guess initial μ_0, σ_0^2

② Forever run:

$$\nabla J \approx \frac{1}{N} \sum_{i=0}^{N} \nabla \log(N(\theta|\mu,\sigma^2)) \sum_{s,a,s' \ldots \in \tau_i} R(s,a,s' \ldots)$$

updates:

$$\mu = \mu + \alpha \nabla_{\mu} J \qquad \sigma^2 = \sigma^2 + \alpha \nabla_{\sigma^2} J$$

$$\text{argmax} \; \underset{N(\theta|\mu,\sigma^2)}{E} R = \text{argmax} \; \underset{N(\theta|\mu,\sigma^2)}{E} \frac{R - \text{mean}}{\text{std}}$$

zero Reward $\quad A = \dfrac{R - E(R)}{Var(R)}$

Scanned by CamScanner

Bellman equations:

state value function $v(s)$

return conditional on state.

$$E[G \mid S_t]$$

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

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

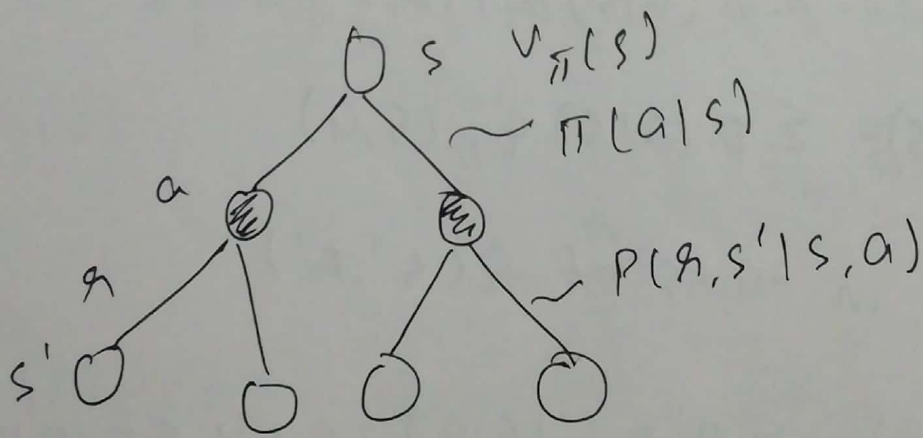
Stochasticity in Policy and enviroment

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

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

Bellman expectation equation for $v(s)$

$$\boxed{V_\pi(s) = E_\pi[R_t + \gamma V_\pi(S_{t+1}) \mid S_t = s]}$$



Backup Daigram

Action value function:

expected return conditional on state & action.

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

no Policy stochasticity at first step.

$$q_\pi(s,a) = E_\pi[R_t + \gamma G_{t+1} \mid S_t = s, A_t = a]$$

$$q_\pi(s,a) = \sum P(r,s' \mid s,a)[r + \gamma E[G_{t+1} \mid S_{t+1} = s']]$$

$$q_\pi(s,a) = \sum_{r,s'} P(r,s' \mid s,a)[r + \gamma V_\pi(S_{t+1})]$$

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

V_π in terms of q_π

$$V_\pi(s) = \sum \pi(a \mid s) \sum P(r,s' \mid s,a)[r + \gamma V(s')]$$

$$V_\pi(s) = \sum \pi(a \mid s)\, q_\pi(s,a)$$

$q(s,a)$ in terms of $q(s',a')$

$$q_\pi(s,a) = \sum p(r,s' \mid s,a)[r + \gamma \sum \pi(a' \mid s')\, q_\pi(s',a')]$$

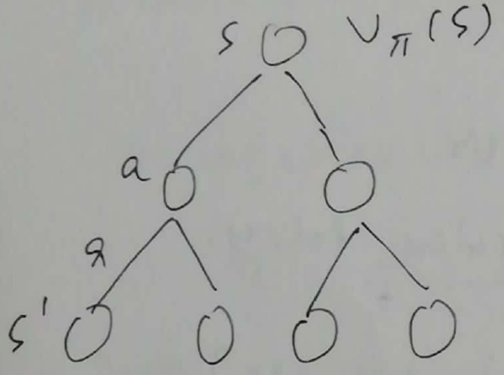
Comparission of Policies: π, π'

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

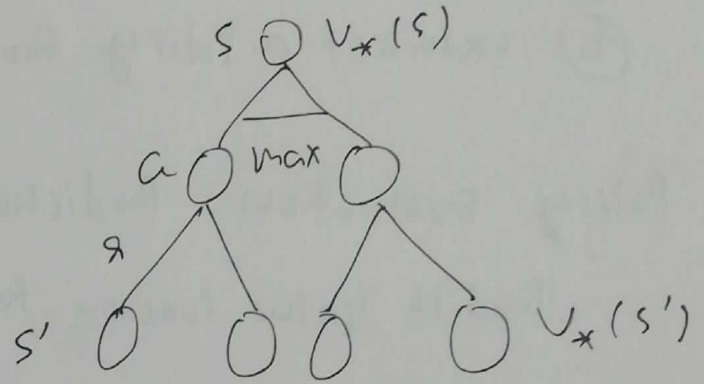
In an MDP atleast 1 π exist satisfying above condition. i.e

In any finite MDP there is always atleast one deterministic optimal policy.

$$v_*(s) = \max_\pi v_\pi(s) \qquad q_*(s,a) = \max_\pi q_\pi(s,a)$$



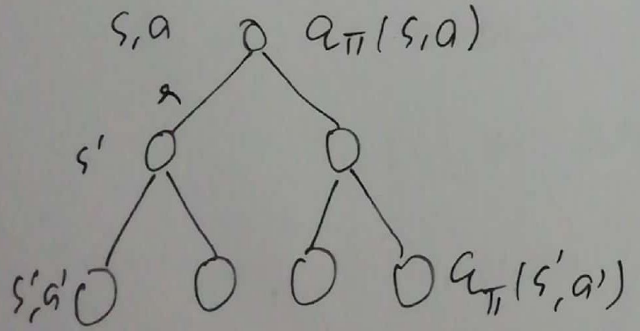
Bellman expectation equation

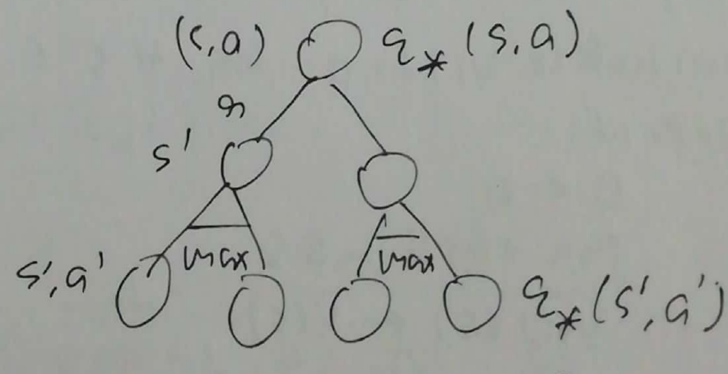


Bellman optimality equation for $v_*(s)$

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

$$= \max_a E_\pi\left[R_t + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a\right]$$





$$q_*(s,a) = E_\pi\left[R_t + \gamma \max_{a'} q_*(S_{t+1}, a') \mid S_t = s, A_t = a\right]$$

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

How to find optimal Policy?

Model & value based.

Model based setup:

Model of world is known ie

$P(r, s' | s, a)$ for all r, s', s, a is known.

Value based setup:

1. Build or estimate a value
2. extract a policy from value.

Policy Evaluation: Prediction problem.

Predict value function for a particular policy.

$$V_\pi(s) = \sum \pi(a|s) \sum P(r, s'|s,a)[r + \gamma V_\pi(s')]$$

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

algorithm iterative:

Input π, policy to be evaluated.

Initialize $V(s) = 0$ for $\forall s \in S$.

Repeat:

$\Delta \leftarrow 0$

For each $s \in S$

$V_{old}(s) \leftarrow V(s)$

$V(s) = \sum \pi(a|s) \cdot \sum P(r, s'|s,a)[r + \gamma V_\pi(s')]$

$\Delta \leftarrow \max(\Delta, |V_{old}(s) - V(s)|)$

until $\Delta < \theta$ (a small positive no.)

Policy improvement :

$$\pi'(s) \leftarrow \arg\max_a \underbrace{\sum P(r, s' | s, a) [r + \gamma V_\pi(s')]}_{q_\pi(s,a)}$$

This procedure is guaranteed to produce better policy.

If $q_\pi(s, \pi'(s)) \geq V_\pi(s)$ for all states

Then $V_{\pi'}(s) \geq V_\pi(s)$

Meaning $\pi' \geq \pi$

Convergence:

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

Then it is optimal.

$$V_{\pi'}(s) = \max_a \sum P(r, s' | s, a) [r + \gamma V_\pi(s')]$$

Determining optimal policy from $V_*(s)$, $q_*(s,a)$

If q_* is known

$$\pi_*(s) \leftarrow \arg\max_a q_*(s,a)$$

If V_* is known

$$\pi_*(s) \leftarrow \arg\max_a \overbrace{\sum P(r, s | s, a) [r + \gamma V_*(s')]}^{q_*(s,a)}$$

In Model free setup we don't know transition probabilities hence we can't convert $V_*(s)$ to $q_*(s)$ so we can't extract π from $V_*(s)$ so we can extract it using $q_*(s)$ only.

Q. Why precise solution to Bellman equation is not needed and approximation is enough?

Generalized Policy Iteration : Policy & Value iteration.

① Evaluate given Policy
② Improve policy by acting greedily wrt value function.

Policy Iteration :

① Evaluate Policy untill convergence (with some tolerance)
② Improve policy.

Value Iteration :

① Evaluate Policy only with single iteration.
② Improve policy.

Policy iteration : scheme

1. Initialize $V(s)$ and $\pi(s)$ $\forall s \in S$.
2. Perform Policy evaluation (without internal initialization).
3. Policy improvement.

Pseudo code :

Policy_stable ← True

for each $s \in S$:

old_action ← $\pi(s)$

$$\pi(s) \leftarrow \arg\max_a \sum \overbrace{P(r, s' \mid s, a)[r + \gamma V(s')]}^{q(s,a)}$$

If old_action $\neq \pi(s)$ then Policy_stable ← false.

If Policy_stable then stop, return $V \approx V_*$, $\pi \approx \pi_*$;

else go to 2.

Value iteration : Scheme.

Initialize V arbitrarily $V(s) = 0$ for all $s \in S$.

Repeat

$\Delta \leftarrow 0$

for each $s \in S$.

$V_{old} \leftarrow V(s)$

{ Bellman optimality equation }

$$V(s) \leftarrow \max_a \sum \overbrace{P(r, s' \mid s, a)[r + \gamma V(s')]}^{q(s,a)}$$

$$\Delta \leftarrow \max(\Delta, |V_{old}(s) - V(s)|)$$

until $\Delta < \theta$

output deterministic Policy $\pi \approx \pi_*$

$$\pi(s) = \arg\max_a \sum_{s', r} P(r, s' \mid s, a)[r + \gamma V(s')]$$

intermediate V(s) in VI might not correspond to any real Policy as its just approximation towards evaluation.

Value Iteration (VI) v/s Policy Iteration (PI)

VI is fast per cycle	PI is slower per cycle
$O(\lvert A\rvert \lvert S\rvert^2)$	$O(\lvert A\rvert \lvert S\rvert^2 + \lvert S\rvert^3)$
VI requires many cycle.	PI req. few cycle.

Model free Policies:

1. Monte-Carlo:

   Get all trajectories containing particular (s, a)
   Estimate G(s, a) for each trajectory.
   Average them to get expectation

Asynchronous Dynamic programming:

in-place iterative DP algo's.

It must compute value of all the states.

Policy evaluation and policy improvement is interleaved.

like:

loop:

$\Delta \leftarrow 0$

loop: ~~for s ∈ S~~ until every state is evaluated.

$v_{old} \leftarrow V(s)$

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

$$\pi(s) = \operatorname*{argmax}_{a} \sum P(r, s' \mid s, a)\left[r + \gamma V(s')\right]$$

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

$\cancel{\text{until } \Delta < \theta}$

Generalized Policy Iteration: