MRP is a tuple (S, P, R, γ), where S is the set of states P is 2-dimensional transition matrix R is a reward function defined as $R = E[r_{t+1}|S_t = s]$, that is collected upon leaving a the state s, i.e. at time step $t+1$ γ is the discount factor
Return - Total discounted reward at from time step t

$R_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... = \sum_{k=0}^{n} \gamma^k R_{t+k+1}$ - γ close to 0 leads to short-sighted evaluation (i.e. don't look into the future too much) - γ close to 1 leads to far-sighted evaluation This return is specific for some sample of states. It does not take into account all possible paths that can be taken from a given state

Value

The aggregated return for a given state can be thought of as its 'value'

$v(s) = E[R_t|S_t = s]$

Bellman Equation for MRP

$$ \begin{align*} v(s) &= E[R_t|S_t=s] \\ &= E[r_{t+1} + \gamma r_{t+2} + \gamma^2r_{t+3}+\cdots|S_t=s] \\ &= E[r_{t+1} + \gamma(r_{t+2} + \gamma r_{t+3}+\cdots)|S_t=s] \\ &= E[r_{t+1} + \gamma R_{t+1}|S_t=s] \tag{1} \\ &= E[r_{t+1} + \gamma E[R_{t+1}|S_t]|S_t=s] \tag{2} \\ &= E[r_{t+1} + \gamma v(s_{t+1})|S_t=s] \end{align*} $$

The transition from (1) to (2) uses the Tower rule and in essence the conditional in the inner expected value will be averaged out/canceled out so it's value does not matter. Therefore it is more useful to use the value of the next state s_{t+1}

See proof: https://chat.openai.com/share/77adb96c-1d17-4fca-8d88-91792567c997

Different Notation

Sum notation (written out)

$$v(s) = R_s + \gamma \sum P_{ss'} v(s')$$

Vector form

$$v = R + \gamma P v$$

This can be analytically solved: $$

$$\begin{aligned} v &= R + \gamma P v \\ v - \gamma P v &= R \\ v(1 - \gamma P) &= R \\ v &= (1 - \gamma P)^{-1} R \end{aligned}$$

$$ Matrix inversion however has a complexity of $O(n^3)$ and therefore works only for small MRPs

Policy (π)

MRP helps us observe the state space and figure out the value of each state. In order to actually influence the environment and take actions a policy π is introduced.

A policy could either be deterministic and non-deterministic.

State Action Value

In order to give a certain value of taking some action, we can define a metric over a given policy π.

$$Q^{\pi}(s, a) = E[R_t|A = a, S_t = s] = E[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}|A = a, S_t = s]$$

In other words we evaluate the return when we take an action a and we follow the policy π thereafter.

A relationship between the state action value and the value of a particular state can be established quite intuitively:

$$V(s) = \sum_{a \in A} \pi(s, a) Q^{\pi}(s, a)$$

Optimal Values Value functions define ordering over policies. A policy is defined to be better or equal to another policy if it's expected return is greater or equal to the one of the other policy for all states s.

$$\pi \geq \pi' \Leftrightarrow V^{\pi}(s) \geq V^{\pi'}(s), \ \forall s \in S$$

Optimal State Value

From this definition we can define the optimal value V^* of a state s using the maximal returns from the optimal policy:

$$V^*(S) = \max_{\pi} V^{\pi}(s),\ \forall s \in S$$

Optimal Policy Analogously the optimal policy being is defined as:

$$\pi^* = \arg\max_{\pi} V^{\pi}(s),\ \forall s \in S$$

Optimal Action Value

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a),\ \ \forall s \in S, \forall a \in A$$

It can also be expressed using the optimal state value:

$$Q^*(s,a) = E[r_{t+1} + \gamma V^*(s_{t+1}) | S = s, A = a]$$

Bellman Optimality for a State Value

The state value can be expressed by finding the action that yields the highest returns. This way the policy is defined and the optimality does not need to refer to any specific policy π. This is useful because policies are often times hard to derive.

$$V^*(s) = \max_{a \in A} \sum_{a \in A} \pi(s,a) P(s'|s,a) Q(s,a)$$

As stated above the policy is already predefined to select the action that returns the greatest value, the policy can be omitted:

$$V^* = \max_{a \in A} \sum_{a \in A} P(s'|s,a)(r(s,a,s') + \gamma V^*(s'))$$

Similarly the state action value can be expressed:

$$\begin{aligned} Q^*(a,s) &= E[r_{t+1} + \max_{a' \in A} \gamma Q^*(s',a') | S = s, A = a] \\ &= \sum P(s,a,s')(r_{t+1} + \max_{a' \in A} \gamma Q^*(s',a')) \end{aligned}$$