MRP is a tuple $(S, A, P, R, \gamma, \pi)$, where S is the state space A is the action space $P^a_{ss'}$ is the transition probability $p(s_{t+1}|a,s_t)$ $R^a_{ss'}$ is a reward function defined as $r_s=r(s,a,s')=r(s_t, s_{t+1},a)$, that is collected upon leaving a the state s, i.e. at time step $t+1$ γ is the discount factor π policy which determines what action should be taken $\pi(a|s)$ - Stochastic: the action is distributed over the policy $a \sim \pi(a|s)$ - Deterministic: the action is set to the result of the policy $a=\pi(a|s)$

Value of a state

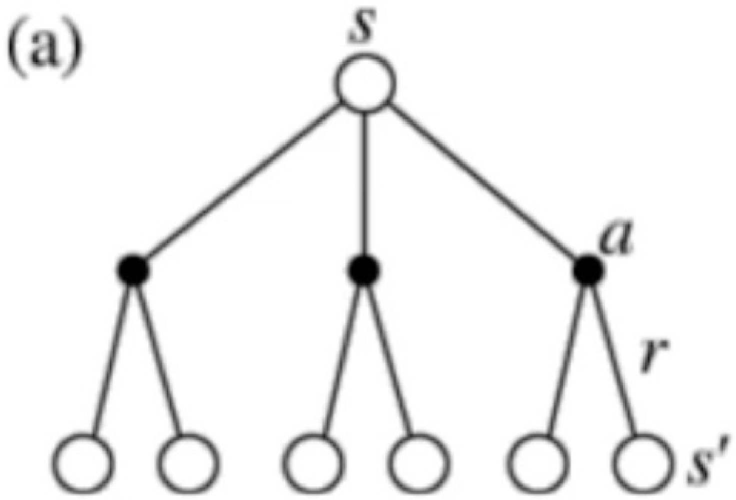
The value of a state V^π given a certain policy π will be followed thereafter, is the expected discounted return after following the policy: $$\begin{align*} V^\pi(s)=\mathbb{E}_{\pi}[R_t|S_t=s] &=\mathbb{E}_{\pi}\bigg[\sum\limits_{k=0}^\infty \gamma^kr_{t+k+1}\bigg|S_t=s\bigg] \\&= \mathbb{E_\pi}\bigg[r_{t+1}+\sum\limits_{k=1}^\infty \gamma^{k}r_{t+k+1}\bigg|S_t=s\bigg] \\&= \mathbb{E_\pi}\bigg[r_{t+1}+\gamma\sum\limits_{k=1}^\infty \gamma^{k-1}r_{t+k+1}\bigg|S_t=s\bigg] \\&= \mathbb{E_\pi}\bigg[r_{t+1}+\gamma\sum\limits_{k=0}^\infty \gamma^{k}r_{t+k+2}\bigg|S_t=s\bigg] \\&= \mathbb{E_\pi}\big[r_{t+1}|S_t=s\big]\tag{1} \\&+ \gamma\mathbb{E_\pi}\bigg[\sum\limits_{k=0}^\infty \gamma^{k}r_{t+k+2}\bigg|S_t=s\bigg]\tag{2} \end{align*}$$

Equation (1)

$$\begin{align*} \mathbb{E_\pi}\big[r_{t+1}|S_t=s\big] \end{align*}$$

The first equation considers the expected immediate award given we are in state s. In other words the question is. ==Note the difference between the return R_t and the immediate reward r_t==

The immediate reward can be easily calculated using a backup tree like this one:



- The root of the tree denotes the current state s.
- The tree branches off into every possible state s' that we can end up at from state s *The question now is what is the probability to take a certain action?*
- This is given by the policy $\pi(a|s)$

Next question is what is the probability that we end up in a state s' given an action a is taken? (The model might perform an action, but the action might not succeed 100% of the time, e.g. try to move left, but the environment pushed the agent forward instead) - This is given by the transition probability matrix $P^a_{ss'}=P(s,a,s')$

Therefore equation (1) could be expressed as:

$$\begin{align*} \mathbb{E_\pi}\big[r_{t+1}|S_t=s\big]&=\sum\limits_{a\in A}\pi(a|s)\sum\limits_{s'\in S} P(s'|s,a)r(s,a,s') \end{align*}$$ ==The expression uses the immediate when moving to the next state. This part of the equation is concerned only with what reward does the agent get rather than the value of the state==

Equation (2)

$$\gamma\mathbb{E_\pi}\bigg[\sum\limits_{k=0}^\infty \gamma^{k}r_{t+k+2}\bigg|S_t=s\bigg]$$ - The γ has been moved outside as the expected value does not change by constants - This equation looks very much like the value of the state s', but the difference is that it is conditioned on $S_t=s$ ***and not*** on $S_t=s'$ like shown below: $$V^\pi(s')=\mathbb{E_\pi}\bigg[\sum\limits_{k=0}^\infty \gamma^{k}r_{t+k+2}\bigg|S_t=s'\bigg]$$ - In order to calculate (2) a similar approach needs to be taken as in equation (1) with the backup tree incorporating the dynamics of the environment: the policy $\pi(a|s)$ and the transition probability matrix $P(s'|a,s)$ $$\begin{align*} \gamma\mathbb{E_\pi}\bigg[\sum\limits_{k=0}^\infty \gamma^{k}r_{t+k+2}\bigg|S_t=s\bigg]&=\sum\limits_{a\in A}\pi(a|s)\sum\limits_{s'\in S} P(s'|s,a)\gamma\sum\limits_{k=0}^\infty \gamma^{k}r_{t+k+2} \\\\&=\sum\limits_{a\in A}\pi(a|s)\sum\limits_{s'\in S} P(s'|s,a)\gamma V^\pi(s') \end{align*}$$ - After incorporating those, the last term $\sum\limits_{k=0}^\infty \gamma^{k}r_{t+k+2}$, denotes the expected discounted return from state $t+1$, i.e. $V^\pi(s')$ #### Combining the two $$ \[\begin{align*} V^\pi(s)&=\mathbb{E}_{\pi}[R_t|S_t=s] \\&= \mathbb{E_\pi}\bigg[r_{t+1}+\gamma\sum\limits_{k=1}^\infty \gamma^{k-1}r_{t+k+1}\bigg|S_t=s\bigg] \\&=\sum\limits_{a\in A}\pi(a|s)\sum\limits_{s'\in S} P(s'|s,a)\bigg(r(s,a,s')+\gamma\sum\limits_{k=1}^\infty \gamma^{k-1}r_{t+k+1}\bigg) \\\\&=\sum\limits_{a\in A}\pi(a|s)\sum\limits_{s'\in

S}P^a_{ss'}\big(R^a_{ss'}+\gamma V(s')\big) \end{align*}\] $$

Policy Evaluation

- The value function $V(s)$ creates an ordering of all the policies π.
- An optimal policy always exists (sometimes >1)
- An optimal policy needs to yield returns that are greater or equal to any other returns generated by a different policy. In other words - the policy would find the best solution no matter from the starting state $$V^*(s)=\max\limits_\pi V^\pi(s),\forall s \in S$$