

Given a

specification  $R$  of immediate rewards after particular actions

calculate the return  $Q$  of particular actions over time via

$$Q = \lim_{n \rightarrow \infty} Q_n$$

## A generalisation

Given a

specification  $R$  of immediate rewards after particular actions

calculate the return  $Q$  of particular actions over time via

$$Q = \lim_{n \rightarrow \infty} Q_n$$

$$Q_{n+1}(s, s') := R(s, s') + \frac{1}{2} \max\{Q_n(s', s'') \mid \text{arc}=(s', s'')\} \quad (1)$$

$\rightsquigarrow$

$$\begin{aligned} Q_{n+1}(s, a) \approx & \alpha [R(s, a) + \gamma \max\{Q_n(s', a') \mid a' \in A\}] \\ & + (1 - \alpha) Q_n(s, a) \end{aligned} \quad (2)$$

## A generalisation

Given a

specification  $R$  of immediate rewards after particular actions

calculate the return  $Q$  of particular actions over time via

$$Q = \lim_{n \rightarrow \infty} Q_n$$

$$Q_{n+1}(s, s') := R(s, s') + \frac{1}{2} \max\{Q_n(s', s'') \mid \text{arc}=(s', s'')\} \quad (1)$$

$\rightsquigarrow$

$$\begin{aligned} Q_{n+1}(s, a) \approx & \alpha [R(s, a) + \gamma \max\{Q_n(s', a') \mid a' \in A\}] \\ & + (1 - \alpha) Q_n(s, a) \end{aligned} \quad (2)$$

(1)	(2)	
$s'$	$a$	(1) is (2) with action $a$ resulting in $s'$
$1$	$\alpha$	deterministically for $\alpha = 1$ , with $\gamma = \frac{1}{2}$
$\frac{1}{2}$	$\gamma$	$s'$ is learned from experience (environment)

# Markov decision process (MDP)

a 5-tuple  $\langle S, A, p, r, \gamma \rangle$  consisting of

- ▶ a finite set  $S$  of states  $s, s', \dots$
- ▶ a finite set  $A$  of actions  $a, \dots$
- ▶ a function  $p : S \times A \times S \rightarrow [0, 1]$

$p(s, a, s') = \text{prob}(s'|s, a) =$  how probable is  $s'$  after doing  $a$  at  $s$

$$\sum_{s'} p(s, a, s') = 1 \text{ for all } a \in A, s \in S$$

# Markov decision process (MDP)

a 5-tuple  $\langle S, A, p, r, \gamma \rangle$  consisting of

- ▶ a finite set  $S$  of states  $s, s', \dots$
- ▶ a finite set  $A$  of actions  $a, \dots$
- ▶ a function  $p : S \times A \times S \rightarrow [0, 1]$

$p(s, a, s') = \text{prob}(s'|s, a) =$  how probable is  $s'$  after doing  $a$  at  $s$

$$\sum_{s'} p(s, a, s') = 1 \text{ for all } a \in A, s \in S$$

- ▶ a function  $r : S \times A \times S \rightarrow \mathbb{R}$

$r(s, a, s') =$  immediate reward at  $s'$  after  $a$  is done at  $s$

# Markov decision process (MDP)

a 5-tuple  $\langle S, A, p, r, \gamma \rangle$  consisting of

- ▶ a finite set  $S$  of states  $s, s', \dots$
- ▶ a finite set  $A$  of actions  $a, \dots$
- ▶ a function  $p : S \times A \times S \rightarrow [0, 1]$

$p(s, a, s') = \text{prob}(s'|s, a) =$  how probable is  $s'$  after doing  $a$  at  $s$

$$\sum_{s'} p(s, a, s') = 1 \text{ for all } a \in A, s \in S$$

- ▶ a function  $r : S \times A \times S \rightarrow \mathbb{R}$

$r(s, a, s') =$  immediate reward at  $s'$  after  $a$  is done at  $s$

- ▶ a discount factor  $\gamma \in [0, 1]$

# Markov decision process (MDP)

a 5-tuple  $\langle S, A, p, r, \gamma \rangle$  consisting of

- ▶ a finite set  $S$  of states  $s, s', \dots$
- ▶ a finite set  $A$  of actions  $a, \dots$
- ▶ a function  $p : S \times A \times S \rightarrow [0, 1]$

$p(s, a, s') = \text{prob}(s'|s, a) =$  how probable is  $s'$  after doing  $a$  at  $s$

$$\sum_{s'} p(s, a, s') = 1 \text{ for all } a \in A, s \in S$$

- ▶ a function  $r : S \times A \times S \rightarrow \mathbb{R}$

$r(s, a, s') =$  immediate reward at  $s'$  after  $a$  is done at  $s$

- ▶ a discount factor  $\gamma \in [0, 1]$

Missing: policy  $\pi : S \rightarrow A$  (what to do at  $s$ )

## Exercise (Poole & Mackworth, chap 9)

Sam is either fit or unfit

$$S = \{\text{fit}, \text{unfit}\}$$

and has to decide whether to exercise or relax

$$A = \{\text{exercise}, \text{relax}\}.$$



## Exercise (Poole & Mackworth, chap 9)

Sam is either fit or unfit

$$S = \{\text{fit}, \text{unfit}\}$$

and has to decide whether to exercise or relax

$$A = \{\text{exercise}, \text{relax}\}.$$

$p(s, a, s')$  and  $r(s, a, s')$  are  $a$ -table entries for row  $s$ , col  $s'$

exercise	fit      unfit
fit	.99, 8
unfit	.2, 0

relax	fit      unfit
fit	.7, 10
unfit	0, 5

immediate rewards do not  
depend on the resulting state

## Exercise (Poole & Mackworth, chap 9)

Sam is either fit or unfit

$$S = \{\text{fit}, \text{unfit}\}$$

and has to decide whether to exercise or relax

$$A = \{\text{exercise}, \text{relax}\}.$$

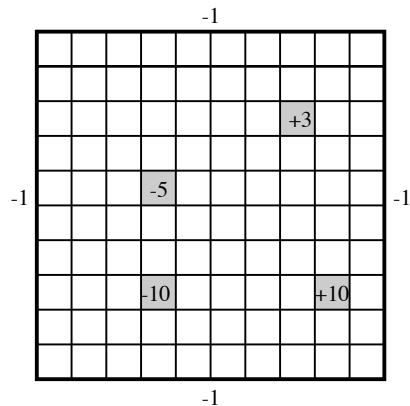
$p(s, a, s')$  and  $r(s, a, s')$  are  $a$ -table entries for row  $s$ , col  $s'$

exercise	fit	unfit	relax	fit	unfit
fit	.99, 8	.01, 8	fit	.7, 10	.3, 10
unfit	.2, 0	.8, 0	unfit	0, 5	1, 5

Entries in red follow from assuming immediate rewards do not depend on the resulting state, and

$$\sum_{s'} p(s, a, s') = 1$$

# Grid World



Poole & Mackworth, 9.5

states: 100 positions  
actions: up, down, left, right  
punish -1 when banging into wall  
& 4 reward/punish states  
prob: 0.7 as directed (if possible)  
...

## Policy from an MDP

Given state  $s$ , pick action  $a$  that maximizes return

$$Q(s, a) := \sum_{s'} \overbrace{p(s, a, s')}^{\text{different outcomes } s'} \left( \underbrace{r(s, a, s')}_{\text{immediate}} + \overbrace{\gamma V(s')}^{\text{discounted future}} \right)$$

for  $V$  tied back to  $Q$  via policy  $\pi : S \rightarrow A$

$$V_{\pi}(s) := Q(s, \pi(s))$$

## Policy from an MDP

Given state  $s$ , pick action  $a$  that maximizes return

$$Q(s, a) := \sum_{s'} \overbrace{p(s, a, s')}^{\text{different outcomes } s'} \left( \underbrace{r(s, a, s')}_{\text{immediate}} + \underbrace{\gamma V(s')}_{\text{discounted future}} \right)$$

for  $V$  tied back to  $Q$  via policy  $\pi : S \rightarrow A$

$$V_{\pi}(s) := Q(s, \pi(s))$$

e.g., the greedy  $Q$ -policy above

$$\pi(s) := \arg \max_a Q(s, a)$$

for

$$Q(s, a) = \sum_{s'} p(s, a, s') (r(s, a, s') + \gamma \max_{a'} Q(s', a'))$$

## Value iteration

Mutual recursion between  $Q/V$  and  $\pi$

value of an action/state vs what to do at a state

## Value iteration

Mutual recursion between  $Q/V$  and  $\pi$

value of an action/state vs what to do at a state

Focus on  $Q$ , approached in the limit

$$\lim_{n \rightarrow \infty} q_n$$

from iterates

$$q_0(s, a) := \sum_{s'} p(s, a, s') r(s, a, s')$$

$$q_{n+1}(s, a) := \sum_{s'} p(s, a, s') (r(s, a, s') + \gamma \max_{a'} q_n(s', a'))$$

## Value iteration

Mutual recursion between  $Q/V$  and  $\pi$

value of an action/state vs what to do at a state

Focus on  $Q$ , approached in the limit

$$\lim_{n \rightarrow \infty} q_n$$

from iterates

$$q_0(s, a) := \sum_{s'} p(s, a, s') r(s, a, s')$$

$$q_{n+1}(s, a) := \sum_{s'} p(s, a, s') (r(s, a, s') + \gamma \max_{a'} q_n(s', a'))$$

In case  $p(s, a, s') = 1$  for some  $s'$  (necessarily unique),  
the iterates simplify to

$$q_0(s, a) := r(s, a, s')$$

$$q_{n+1}(s, a) := r(s, a, s') + \gamma \max_{a'} q_n(s', a')$$



## Deterministic actions and absorbing states (game over)

Fix an MDP with min reward  $m$ .

An action  $a$  is *s-deterministic* if  $p(s, a, s') = 1$  for some  $s'$ .

## Deterministic actions and absorbing states (game over)

Fix an MDP with min reward  $m$ .

An action  $a$  is *s-deterministic* if  $p(s, a, s') = 1$  for some  $s'$ .

A state  $s$  is *absorbing* if  $p(s, a, s) = 1$  for every action  $a$ , whence

$$Q(s, a) = r(s, a, s) + \gamma V(s)$$

$$V(s) = \frac{r_s}{1 - \gamma} \quad \text{where } r_s = \max_a r(s, a, s)$$

A state  $s$  is a *sink* if it is absorbing and  $r(s, a, s) = m$  for all  $a$ .

## Deterministic actions and absorbing states (game over)

Fix an MDP with min reward  $m$ .

An action  $a$  is *s-deterministic* if  $p(s, a, s') = 1$  for some  $s'$ .

A state  $s$  is *absorbing* if  $p(s, a, s) = 1$  for every action  $a$ , whence

$$Q(s, a) = r(s, a, s) + \gamma V(s)$$

$$V(s) = \frac{r_s}{1 - \gamma} \quad \text{where } r_s = \max_a r(s, a, s)$$

A state  $s$  is a *sink* if it is absorbing and  $r(s, a, s) = m$  for all  $a$ .

An action  $a$  is an *s-drain* if for some sink  $s'$ ,

$$p(s, a, s') = 1 \text{ and } r(s, a, s') = m$$

## Deterministic actions and absorbing states (game over)

Fix an MDP with min reward  $m$ .

An action  $a$  is *s-deterministic* if  $p(s, a, s') = 1$  for some  $s'$ .

A state  $s$  is *absorbing* if  $p(s, a, s) = 1$  for every action  $a$ , whence

$$Q(s, a) = r(s, a, s) + \gamma V(s)$$
$$V(s) = \frac{r_s}{1 - \gamma} \quad \text{where } r_s = \max_a r(s, a, s)$$

A state  $s$  is a *sink* if it is absorbing and  $r(s, a, s) = m$  for all  $a$ .

An action  $a$  is an *s-drain* if for some sink  $s'$ ,

$$p(s, a, s') = 1 \text{ and } r(s, a, s') = m$$

Let

$$A(s) := \{a \in A \mid a \text{ is not an } s\text{-drain}\}$$

so if  $A(s) \neq \emptyset$ ,

$$V(s) = \max\{Q(s, a) \mid a \in A\} = \max\{Q(s, a) \mid a \in A(s)\}$$

## Arcs & goals as a deterministic MDP ( $p \in \{0, 1\}$ )

Given *arc* and goal set *G*, let

$$A = \{s \mid (\exists s') \text{ arc}=(s', s)\} = S$$

where for each  $a \in A$ ,

$$p(s, a, s') = \begin{cases} 1 & \text{if } a = s' \text{ and } \text{arc}=(s, s') \\ 0 & \text{otherwise} \end{cases}$$

$$r(s, a, s') = \begin{cases} R(s, s') & \text{if } a = s' \text{ and } \text{arc}=(s, s') \\ \text{anything} & \text{otherwise} \end{cases}$$

## Arcs & goals as a deterministic MDP ( $p \in \{0, 1\}$ )

Given *arc* and goal set  $G$ , let

$$A = \{s \mid (\exists s') \text{ arc}=(s', s)\} = S$$

where for each  $a \in A$ ,

$$p(s, a, s') = \begin{cases} 1 & \text{if } a = s' \text{ and } \text{arc}=(s, s') \\ 0 & \text{otherwise} \end{cases}$$

$$r(s, a, s') = \begin{cases} R(s, s') & \text{if } a = s' \text{ and } \text{arc}=(s, s') \\ \text{anything} & \text{otherwise} \end{cases}$$

Satisfy prob constraint  $\sum_{s'} p(s, a, s') = 1$  via sink state  $\perp \notin A$ , requiring of every  $a \in A$  and  $s \in S$ ,

$$p(s, a, \perp) = \begin{cases} 1 & \text{if not } \text{arc}=(s, a) \\ 0 & \text{otherwise} \end{cases}$$

$$p(\perp, a, s) = \begin{cases} 1 & \text{if } s = \perp \\ 0 & \text{otherwise} \end{cases}$$

$$r(s, a, \perp) = \text{min reward}$$