Given a

specification R of immediate rewards after particular actions

calculate the return Q of particular actions over time via

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

A generalisation

Given a

specification R of immediate rewards after particular actions

calculate the return Q of particular actions over time via

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

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

$\rightsquigarrow$

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

A generalisation

Given a

specification R of immediate rewards after particular actions

calculate the return Q of particular actions over time via

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

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

$\rightsquigarrow$

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

(1)	(2)
s'	a
1	α
$\frac{1}{2}$	γ

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

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 $s, s', \ldots$
- a finite set A of actions $a, \ldots$
- a function $p : S \times A \times S \to [0, 1]$

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

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

Markov decision process (MDP)

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

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

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

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

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

$$r(s, a, s') \ = \ \text{expected immediate reward for } s \xrightarrow{a} s'$$

Markov decision process (MDP)

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

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

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

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

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

$$r(s,a,s') \;=\; \text{expected immediate reward for } s \xrightarrow{a} s'$$

- 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 $s, s', \ldots$
- a finite set A of actions $a, \ldots$
- a function $p : S \times A \times S \to [0, 1]$

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

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

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

$$r(s, a, s') \;=\; \text{expected immediate reward for } s \xrightarrow{a} s'$$

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

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

Exercise (Poole & Mackworth, chap 12)

Sam is either fit or unfit

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

and has to decide whether to exercise or relax

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

Exercise (Poole & Mackworth, chap 12)

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 12)

Sam is either fit or unfit

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

and has to decide whether to exercise or relax

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

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

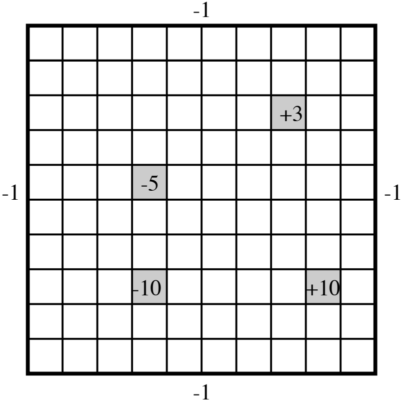
exercise	fit	unfit
fit	.99, 8	.01, 8
unfit	.2, 0	.8, 0

relax	fit	unfit
fit	.7, 10	.3, 10
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, 12.5

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

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

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

$$V_\pi(s) \;:=\; Q(s, \pi(s))$$

Policy from an MDP

Given state s, pick action a that maximizes return

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

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

$$V_\pi(s) \;:=\; Q(s, \pi(s))$$

e.g., the greedy Q-policy above

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

for

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

Value iteration

Mutual recursion between Q/V and π

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

Value iteration

Mutual recursion between Q/V and π

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

Focus on Q, approached in the limit

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

from iterates

$$\begin{aligned} q_0(s,a) &:= \sum_{s'} p(s,a,s') r(s,a,s') \\ q_{n+1}(s,a) &:= \sum_{s'} p(s,a,s') \big(r(s,a,s') + \gamma \max_{a'} q_n(s',a') \big) \end{aligned}$$

Value iteration

Mutual recursion between Q/V and π

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

Focus on Q, approached in the limit

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

from iterates

$$\begin{aligned} q_0(s,a) &:= \sum_{s'} p(s,a,s')r(s,a,s') \\ q_{n+1}(s,a) &:= \sum_{s'} p(s,a,s')\big(r(s,a,s') + \gamma \max_{a'} q_n(s',a')\big) \end{aligned}$$

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

$$\begin{aligned} q_0(s,a) &:= r(s,a,s') \\ q_{n+1}(s,a) &:= r(s,a,s') + \gamma \max_{a'} q_n(s',a') \end{aligned}$$

Determinstic 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'.

Determinstic 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

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

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

Determinstic 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

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

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',

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

Determinstic 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

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

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

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

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

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

where for each $a \in A$,

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

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

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

Given *arc* and goal set G, let

$$A \;=\; \{s \mid (\exists s')\; arc_{=}(s', s)\} \;=\; S$$

where for each $a \in A$,

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

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

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

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

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

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