Reinforcement Learning

Prof. Christian Bauckhage

lecture 08
solving Bellman equations

in this lecture, we have a closer look at some of the concepts which we introduced last time

to begin with, we examine properties of *deterministic policies* (will be useful later)

then, we get to know *Bellman's principle of optimality* which allows us to scrutinize the *Bellman equation* for value functions

crucially, we will see that we are dealing with whole systems of equations and will learn how to *solve systems of Bellman equations*

this will then allow us to perform *policy evaluation*

outline

recap

reinforcement learning

$\Leftrightarrow$ the problem faced by an agent that uses trial-and-error to learn how to behave in a dynamic environment

$\Leftrightarrow$ the problem faced by an agent that uses trial-and-error to learn to decide *what* to do *when* in order to achieve a long-term goal

"the" modeling tool for RL problems are *Markov decision processes* over sets of random variables S_t, A_t, and R_{t+1}

MDP $\Leftrightarrow$ a tuple $(\mathcal{S}, \mathcal{A}, T, R)$ with

set of states

$$\mathcal{S} = \{s_1, s_2, \ldots\}$$

set of actions

$$\mathcal{A} = \{a_1, a_2, \ldots\}$$

transition function / state transition probabilities

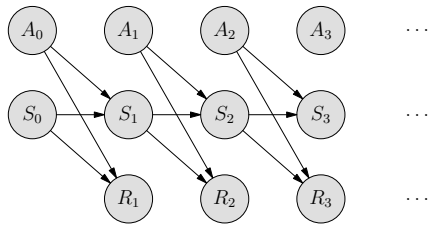
$$T(s', s, a) = p(S_{t+1} = s' \mid S_t = s, A_t = a)$$

reward function / expected immediate rewards

$$R(s, a) = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]$$

MDP $\Leftrightarrow$ a stochastic process over S_t, A_t, and R_t where

$$p\big(S_{t+1} = s', R_{t+1} = r \;\big|\; S_t = s, A_t = a\big) = p\big(s', r \;\big|\; s, a\big)$$



a fully specified MDP $(\mathcal{S}, \mathcal{A}, T, R)$ *models* an environment

an MDP model $(\mathcal{S}, \mathcal{A}, T, R)$ predicts an environment, because

the transition function

$$T(s', s, a) = p(S_{t+1} = s' \mid S_t = s, A_t = a)$$

predicts next states

the reward function

$$R(s, a) = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]$$

predicts next (immediate) rewards

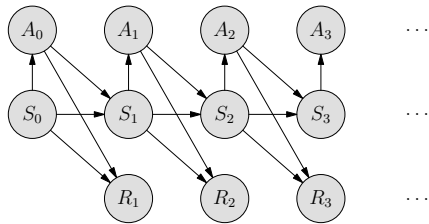
infinite-horizon discounted model of expected returns

$$G_0 = \sum_{t=0}^{\infty} \gamma^t R_{t+1} = \sum_{t=0}^{\infty} \gamma^t R\big(S_t, A_t\big) \quad \text{where} \quad 0 \leqslant \gamma < 1$$

stochastic policy

$$\pi : \mathcal{S} \times \mathcal{A} \to \big[0, 1\big] \quad \text{with} \quad \pi\big(s, a\big) \equiv \pi\big(a \mid s\big) \equiv p\big(a \mid s\big)$$

when introducing / working with a stochastic policy, the MDP model becomes



value function (for a given policy π)

$$V_\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} \middle| S_0 = s\right] = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \middle| S_0 = s\right]$$

quality function (for a given policy π)

$$Q_\pi(s,a) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} \middle| S_0 = s, A_0 = a\right] = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \middle| S_0 = s, A_0 = a\right]$$

Bellman equations

$$V_\pi(s) = \sum_a R(s,a)\,\pi(a \mid s) + \gamma \sum_a \sum_{s'} V_\pi(s')\,p(s' \mid s,a)\,\pi(a \mid s)$$

$$Q_\pi(s,a) = R(s,a) + \gamma \sum_{s'} \sum_{a'} Q_\pi(s',a')\,\pi(a' \mid s')\,p(s' \mid s,a)$$

note the “next state” s' recursions

note

we also have

$$V_\pi(s) = \sum_a Q_\pi(s,a)\,\pi(a \mid s)$$

$$Q_\pi(s,a) = R(s,a) + \gamma \sum_{s'} V_\pi(s')\,p(s' \mid s,a)$$

note

$\Leftrightarrow$ we also have

$$V_\pi(s) = \mathbb{E}_\pi\Big[Q_\pi(S_0, A_0) \;\Big|\; S_0 = s\Big]$$

$$Q_\pi(s, a) = \mathbb{E}_p\Big[R(S_0, A_0) + \gamma\, V_\pi(S_1) \;\Big|\; S_0 = s, A_0 = a\Big]$$

note

these equations are Bellman equations

$$V_\pi(s) = \sum_a R(s,a)\,\pi(a \mid s) + \gamma \sum_a \sum_{s'} V_\pi(s')\,p(s' \mid s,a)\,\pi(a \mid s)$$

$$Q_\pi(s,a) = R(s,a) + \gamma \sum_{s'} \sum_{a'} Q_\pi(s',a')\,\pi(a' \mid s')\,p(s' \mid s,a)$$

because they relate $V_\pi(s)$ to $V_\pi(s')$ and $Q_\pi(s,a)$ to $Q_\pi(s',a')$, respectively

these equations are useful

$$V_\pi(s) = \sum_a Q_\pi(s,a)\,\pi(a \mid s)$$

$$Q_\pi(s,a) = R(s,a) + \gamma \sum_{s'} V_\pi(s')\,p(s' \mid s,a)$$

but they relate $V_\pi(s)$ to $Q_\pi(s,a)$ and $Q_\pi(s,a)$ to $V_\pi(s)$, respectively

note

all of this generalizes to

- continuous time evolution
- continuous state spaces
- continuous action spaces
- partially observable MDPs (POMDPs)

we will study generalizations later on . . .

working with deterministic policies

disclaimer

what we will do next is nothing but a special case of what we did already !

our main reasons for considering this special case are to hit home a point about the reward function $R(s, a)$ and to be able to define the concept of a *Markov reward process*

we therefore recall from lecture 01 ...

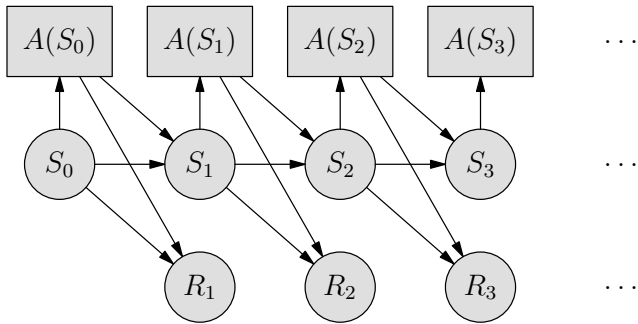
a deterministic policy is a *function*

$$\pi : \mathcal{S} \to \mathcal{A} \quad \text{with} \quad \pi(s) = a$$

which maps a state to an action

some people write $\mu(s)$ instead of $\pi(s)$

Bayesian network representation



note

under a deterministic policy π, an MDP transition function becomes

$$\begin{aligned} T\big(s', s, \pi(s)\big) &= p\big(S_{t+1} = s' \;\big|\; S_t = s, A_t = \pi(s)\big) \\ &= p_\pi\big(S_{t+1} = s' \;\big|\; S_t = s\big) \\ &= T_\pi\big(s', s\big) \end{aligned}$$

under a deterministic policy π, an MDP reward function becomes

$$\begin{aligned} R\big(s, \pi(s)\big) &= \mathbb{E}\big[R_{t+1} \;\big|\; S_t = s, A_t = \pi(s)\big] = \sum_{s'} Util\big(s'\big)\, p_\pi\big(s' \;\big|\; s\big) \\ &= \mathbb{E}_{p_\pi}\big[R_{t+1} \;\big|\; S_t = s\big] \\ &= R_\pi\big(s\big) \end{aligned}$$

question

wait a minute . . .

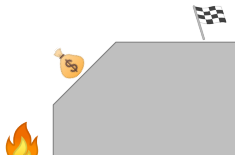
if action selection is deterministic, then why are state transitions still probabilistic ?

answer

the agent's environment may be antagonistic / non-deterministic / non-controllable

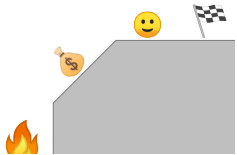
consider this . . .

recall the simple game environment

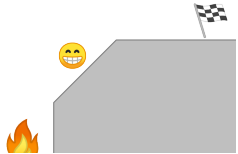


example inspired by Elliot Waite

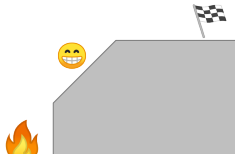
upon spawning, you go left . . .



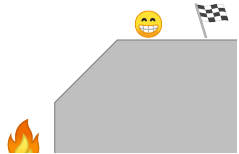
and collect the loot

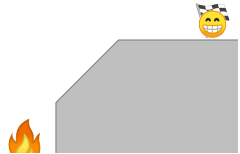


of course you next go right . . .

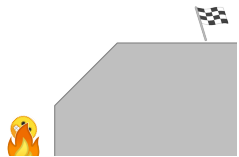


with a 90% chance you succeed and then win





alas, with a 10% chance you slip into the fire



note

$\Rightarrow$ in general, selected actions can't be guaranteed to have the desired effect

this thus explains why is is common to work with *expected rewards*

$$R(s,a) = \mathbb{E}\left[R_{t+1} \mid S_t = s, A_t = a\right] = \sum_{s'} Util(s')\, p(s' \mid s,a)$$

since executing action a in state s may result in various successor states s'

one more thing . . .

everybody writes the reward function as

$$R(s,a) = \mathbb{E}\left[R_{t+1} \mid S_t = s, A_t = a\right]$$

but it would of course be more precise to write

$$\begin{aligned} R(s,a) &= \sum_{s'} Util(s')\, p(s' \mid s,a) \\ &= \mathbb{E}_p\left[R_{t+1} \mid S_t = s, A_t = a\right] \end{aligned}$$

later in this lecture, we will be more precise and write $\mathbb{E}_p\left[\cdot\right]$

note

combining a Markov decision process with a deterministic policy leads to *deterministic action selection* in each state

$\Leftrightarrow$ a Markov decision process combined with a deterministic policy behaves like a Markov chain, because

$$p\big(s' \,\big|\, s, \pi(s)\big) = p_\pi\big(s' \,\big|\, s\big)$$

since there also are outputs $R_\pi(s)$, we are dealing with ...

Markov reward process (MRP)

$\Leftrightarrow$ a Markov chain with values

$\Leftrightarrow$ a tuple $(\mathcal{S}, T, R)$ such that

$$\mathcal{S} \equiv \text{set of states}$$

$$\begin{aligned} T(s', s) &\equiv \text{state transition probabilities} \\ &= p(S_{t+1} = s' \mid S_t = s) \end{aligned}$$

$$\begin{aligned} R(s) &\equiv \text{expected values / rewards} \\ &= \mathbb{E}[R_{t+1} \mid S_t = s] \end{aligned}$$

using this, we may also say ...

Markov decision process (MDP)

$\Leftrightarrow$ a Markov reward process augmented by decisions

Bellman equations revisited

and back to stochastic policies $\pi(a \mid s)$ ...

an optimal value function / optimal value of a state is given by

$$V_*(s) = \max_\pi \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R\Big(S_t, \pi\big(A_t \mid S_t\big)\Big) \middle| S_0 = s \right]$$

$$= \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t R\Big(S_t, \pi_*\big(A_t \mid S_t\big)\Big) \middle| S_0 = s \right]$$

an optimal policy is a policy $\pi_*(a \mid s)$ such that for all states

$$V_*(s) = V_{\pi_*}(s) \geqslant V_\pi(s)$$

note

we seem to be going in circles, since

$V_*(s)$ is defined in terms of $\pi_*(s)$

$\pi_*(s)$ is defined in terms of $V_*(s)$

note

we seem to be going in circles, since

$V_*(s)$ is defined in terms of $\pi_*(s)$

$\pi_*(s)$ is defined in terms of $V_*(s)$

enter Richard Bellman ...

Bellman's principle of optimality

An optimal policy has the property that whatever the initial state and the initial decision are, the remaining decisions must constitute an optimal policy w.r.t. the state resulting from the first decision.

R.E. Bellman
($*$1920, †1984)

Bellman equations

$\Leftrightarrow$ use the principle of optimality to relate a value function to itself

$$V_\pi(s) = \sum_a R(s,a)\,\pi(a \mid s) + \gamma \sum_a \sum_{s'} V_\pi(s')\,p(s' \mid s,a)\,\pi(a \mid s)$$

$$= \mathbb{E}_\pi\left[\mathbb{E}_p\left[R_1 \;\middle|\; S_0 = s, A_0\right] + \gamma\,\mathbb{E}_p\left[V_\pi(S_1) \;\middle|\; S_0 = s, A_0\right]\right]$$

$\Leftrightarrow$ the (expected) value of a state is its immediate return plus the (discounted) average of the values of its possible successors

we have

$$
\begin{aligned}
V_\pi(s) &= \sum_a R(s,a)\,\pi(a \mid s) + \gamma \sum_a \sum_{s'} V_\pi(s')\,p(s' \mid s,a)\,\pi(a \mid s) \\
&= \sum_a \Big[R(s,a) + \gamma \sum_{s'} V_\pi(s')\,p(s' \mid s,a) \Big] \pi(a \mid s) \\
&= \mathbb{E}_\pi \Big[R(s,a) + \gamma \sum_{s'} V_\pi(s')\,p(s' \mid s,a) \Big] \\
&= \mathbb{E}_\pi \Big[\mathbb{E}_p \big[R_1 \;\big|\; S_0 = s, A_0 \big] + \gamma \sum_{s'} V_\pi(s')\,p(s' \mid s,a) \Big] \\
&= \mathbb{E}_\pi \Big[\mathbb{E}_p \big[R_1 \;\big|\; S_0 = s, A_0 \big] + \gamma\, \mathbb{E}_p \big[V_\pi(S_1) \;\big|\; S_0 = s, A_0 \big] \Big] \\
&= \mathbb{E}_\pi \Big[\mathbb{E}_p \big[R_1 \;\big|\; S_0 = s, A_0 \big] \Big] + \gamma\, \mathbb{E}_\pi \Big[\mathbb{E}_p \big[V_\pi(S_1) \;\big|\; S_0 = s, A_0 \big] \Big]
\end{aligned}
$$

from lecture 07, we recall

$$
\mathbb{E}_\pi \Big[\mathbb{E}_p \big[R_1 \mid S_0 = s, A_0 \big] \Big] = \mathbb{E}_p \big[R_1 \mid S_0 = s \big] \equiv R_\pi(s)
$$

recall

$$\mathbb{E}\Big[\mathbb{E}\big[X\big]\Big] = \mathbb{E}\big[X\big]$$

recall

$$\mathbb{E}_\pi\Big[R(s,a) \;\Big|\; S_0 = s\Big] = \mathbb{E}_\pi\Big[\mathbb{E}_p\big[R_1 \mid S_0, A_0\big] \;\Big|\; S_0 = s\Big] = \mathbb{E}_p\big[R_1 \mid S_0 = s\big] \equiv R(s)$$

observe

$$\mathbb{E}_\pi\Big[R(s) \;\Big|\; s\Big] = R(s)$$
$$\mathbb{E}_p\Big[R(s) \;\Big|\; s\Big] = R(s)$$

to reduce clutter, we next simply write

$$\mathbb{E}_\pi\Big[\mathbb{E}_p\Big[R_1 \;\Big|\; S_0, A_0\Big] + \gamma\,\mathbb{E}_p\Big[V_\pi(s') \;\Big|\; S_0, A_0\Big] \;\Big|\; S_0 = s\Big] = R(s) + \gamma\,\mathbb{E}_\pi\Big[\mathbb{E}_p\Big[V_\pi(s')\Big] \;\Big|\; S_0 = s\Big]$$

using *telescoping*, we find

$$
\begin{aligned}
V_\pi(s) &= R_\pi(s) + \gamma\, \mathbb{E}_\pi\Big[\mathbb{E}_p\Big[V_\pi(S_1) \;\Big|\; S_0 = s, A_0\Big]\Big] \\
&= R_\pi(s) + \gamma\, \mathbb{E}_\pi\Big[\mathbb{E}_p\Big[R_\pi(s') + \gamma\, \mathbb{E}_\pi\Big[\mathbb{E}_p\Big[V_\pi(S_2) \;\Big|\; S_1 = s', A_1\Big]\Big]\Big] \\
&= R_\pi(s) + \gamma\, R_\pi(s') + \gamma^2\, \mathbb{E}_\pi\Big[\mathbb{E}_p\Big[V_\pi(S_2) \;\Big|\; S_1 = s', A_1\Big]\Big] \qquad \text{(double expectations "cancel")} \\
&\;\;\vdots \\
&= R_\pi(s) + \gamma\, R_\pi(s') + \gamma^2\, R_\pi(s'') + \gamma^3\, R_\pi(s''') + \cdots \\
&= \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t\, R(S_t, A_t) \,\middle|\, S_0 = s\right]
\end{aligned}
$$

in short, we find

$$
\begin{aligned}
V_\pi(s) &= \sum_a R(s,a)\,\pi(a \mid s) + \gamma \sum_a \sum_{s'} V_\pi(s')\,p(s' \mid s,a)\,\pi(a \mid s) \\
&= \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \,\middle|\, S_0 = s\right] \\
&= \mathbb{E}_\pi\Big[G_0 \;\Big|\; S_0 = s\Big]
\end{aligned}
$$

$\Rightarrow$ Bellman's principle of optimality *implies* the definition of $V_\pi(s)$ we gave earlier

remark

above, we were excruciatingly pedantic and found

$$V_\pi(s) = \mathbb{E}_\pi\Big[G_0 \;\Big|\; S_0 = s\Big]$$
$$= \mathbb{E}_\pi\Big[R_1 + \gamma\, G_1 \;\Big|\; S_0 = s\Big]$$

however, in the literature, it is very common to read

$$V_\pi(s) = \mathbb{E}_\pi\Big[R_1 + \gamma\, V_\pi(S_1) \;\Big|\; S_0 = s\Big]$$

remark (continued)

the equality

$$\mathbb{E}_\pi\Big[R_1 + \gamma\, G_1 \;\Big|\; S_0 = s\Big] = \mathbb{E}_\pi\Big[R_1 + \gamma\, V_\pi(S_1) \;\Big|\; S_0 = s\Big]$$

can be confusing, because we actually have

$$G_1 \neq V_\pi(S_1)$$

we do, however, have *equality in expectation*

$$\mathbb{E}_\pi\Big[G_1 \;\Big|\; S_0 = s\Big] = \mathbb{E}_\pi\Big[V_\pi(S_1) \;\Big|\; S_0 = s\Big]$$

question

but this doesn't involve *optimal* $V_*(s)$ or $\pi_*(s)$, or does it ?

answer

no, it doesn't yet . . . we will get back to this point later . . .

note

there actually is one Bellman equation for each state $s \in \mathcal{S}$

$$V_\pi(s_1) = \sum_a R(s_1, a)\, \pi(a \mid s_1) + \gamma \sum_a \sum_{s'} V_\pi(s')\, p(s' \mid s_1, a)\, \pi(a \mid s_1)$$

$$V_\pi(s_2) = \sum_a R(s_2, a)\, \pi(a \mid s_2) + \gamma \sum_a \sum_{s'} V_\pi(s')\, p(s' \mid s_2, a)\, \pi(a \mid s_2)$$

$$\vdots$$

$\Rightarrow$ **Bellman equations always constitute whole systems of equations**

let us rearrange the Bellman equations as

$$V_\pi(s) = \sum_a R(s,a)\,\pi(a \mid s) + \gamma \sum_a \sum_{s'} V_\pi(s')\,p(s' \mid s,a)\,\pi(a \mid s)$$

$$= \underbrace{\sum_a R(s,a)\,\pi(a \mid s)}_{=T_1} + \gamma \underbrace{\sum_{s'} V_\pi(s') \sum_a p(s' \mid s,a)\,\pi(a \mid s)}_{=T_2}$$

and examine the deeper meaning / consequences of the insight we just had

to begin with, we note ...

if the state set $\mathcal{S}$ is discrete and of finite size $|\mathcal{S}| = n$, we can think of a value function $V_\pi(\cdot)$ as a vector

$$\boldsymbol{v}_\pi = \begin{bmatrix} V_\pi(s_1) \\ V_\pi(s_2) \\ \vdots \end{bmatrix}$$

next, we note ...

if the action set $\mathcal{A}$ is discrete and of finite size $|\mathcal{A}| = m$, we can think of reward functions $R(\cdot, \cdot)$ and stochastic policies $\pi(\cdot, \cdot)$ as matrices

$$\boldsymbol{R} = \begin{bmatrix} R(a_1, s_1) & R(a_1, s_2) & \cdots \\ R(a_2, s_1) & R(a_2, s_2) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

$$\boldsymbol{\Pi} = \begin{bmatrix} \pi(a_1 \mid s_1) & \pi(a_1 \mid s_2) & \cdots \\ \pi(a_2 \mid s_1) & \pi(a_2 \mid s_2) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

next, we note ...

$\Rightarrow$ w.r.t. T_1, we realize that

$$\sum_a R(s,a)\,\pi(a \mid s)$$

computes a vector, namely

$$\boldsymbol{r}_\pi^\intercal = \mathbf{1}^\intercal \left[\boldsymbol{R} \odot \boldsymbol{\Pi}\right]$$

$$\Leftrightarrow \quad \boldsymbol{r}_\pi = \left[\boldsymbol{R} \odot \boldsymbol{\Pi}\right]^\intercal \mathbf{1}$$

next, we note ...

for the inner sum of T_2, we have the following equality

$$\sum_a p(s' \mid s,a) \cdot \pi(a \mid s) = \sum_a \frac{p(s',s,a)}{p(s,a)} \cdot \frac{p(s,a)}{p(s)} = \sum_a \frac{p(s',a \mid s)\, p(s)}{p(s)} = p(s' \mid s)$$

and can arrange all the resulting state transition probabilities $p(s' \mid s)$ in a matrix

$$\boldsymbol{P}_\pi = \begin{bmatrix} p(s_1 \mid s_1) & p(s_1 \mid s_2) & \cdots \\ p(s_2 \mid s_1) & p(s_2 \mid s_2) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

moreover, since this matrix $\boldsymbol{P}_\pi$ is column stochastic, its spectral radius is $\rho(\boldsymbol{P}_\pi) = 1$

$\Rightarrow$ w.r.t. T_2, we realize that

$$\sum_{s'} V_\pi(s') \sum_a p(s' \mid s, a)\, \pi(a \mid s)$$

computes a vector, namely

$$\boldsymbol{u}^\mathsf{T} = \boldsymbol{v}_\pi^\mathsf{T} \boldsymbol{P}_\pi$$

$$\Leftrightarrow \quad \boldsymbol{u} = \boldsymbol{P}_\pi^\mathsf{T} \boldsymbol{v}_\pi$$

note

$\Rightarrow$ we can rewrite a system of Bellman equations
in terms of just a single matrix-vector equation

$$\boldsymbol{v}_\pi = \boldsymbol{r}_\pi + \gamma \boldsymbol{P}_\pi^\mathsf{T} \boldsymbol{v}_\pi$$

we can solve this equation for $\boldsymbol{v}_\pi$ because

$$\boldsymbol{v}_\pi = \boldsymbol{r}_\pi + \gamma\, \boldsymbol{P}_\pi^\intercal\, \boldsymbol{v}_\pi$$

$$\Leftrightarrow \quad \boldsymbol{v}_\pi - \gamma\, \boldsymbol{P}_\pi^\intercal\, \boldsymbol{v}_\pi = \boldsymbol{r}_\pi$$

$$\Leftrightarrow \quad \left[\boldsymbol{I} - \gamma\, \boldsymbol{P}_\pi^\intercal\right] \boldsymbol{v}_\pi = \boldsymbol{r}_\pi$$

$$\Leftrightarrow \quad \boldsymbol{v}_\pi = \left[\boldsymbol{I} - \gamma\, \boldsymbol{P}_\pi^\intercal\right]^{-1} \boldsymbol{r}_\pi \qquad (\gamma < 1 \wedge \rho(\boldsymbol{P}_\pi^\intercal) = 1 \;\Rightarrow\; \rho(\gamma\, \boldsymbol{P}_\pi^\intercal) < 1)$$

now, let $\boldsymbol{v}_0$ be some initial value vector and consider this process

$$\boldsymbol{v}_1 = \boldsymbol{r}_\pi + \gamma \boldsymbol{P}_\pi^\intercal \boldsymbol{v}_0$$

$$\begin{aligned}\boldsymbol{v}_2 &= \boldsymbol{r}_\pi + \gamma \boldsymbol{P}_\pi^\intercal \boldsymbol{v}_1 \\ &= \boldsymbol{r}_\pi + \gamma \boldsymbol{P}_\pi^\intercal \left[\boldsymbol{r}_\pi + \gamma \boldsymbol{P}_\pi^\intercal \boldsymbol{v}_0\right] = \boldsymbol{r}_\pi + \gamma \boldsymbol{P}_\pi^\intercal \boldsymbol{r}_\pi + \gamma^2 {\boldsymbol{P}_\pi^\intercal}^2 \boldsymbol{v}_0\end{aligned}$$

$$\begin{aligned}\boldsymbol{v}_3 &= \boldsymbol{r}_\pi + \gamma \boldsymbol{P}_\pi^\intercal \boldsymbol{v}_2 \\ &= \boldsymbol{r}_\pi + \gamma \boldsymbol{P}_\pi^\intercal \left[\boldsymbol{r}_\pi + \gamma \boldsymbol{P}_\pi^\intercal \boldsymbol{r}_\pi + \gamma^2 {\boldsymbol{P}_\pi^\intercal}^2 \boldsymbol{v}_0\right] = \boldsymbol{r}_\pi + \gamma \boldsymbol{P}_\pi^\intercal \boldsymbol{r}_\pi + \gamma^2 {\boldsymbol{P}_\pi^\intercal}^2 \boldsymbol{r}_\pi + \gamma^3 {\boldsymbol{P}_\pi^\intercal}^3 \boldsymbol{v}_0 \\ &\vdots\end{aligned}$$

$$\boldsymbol{v}_t = \sum_{\tau=0}^{t-1} \left[\gamma \boldsymbol{P}_\pi^\intercal\right]^\tau \boldsymbol{r}_\pi + \left[\gamma \boldsymbol{P}_\pi^\intercal\right]^t \boldsymbol{v}_0 \qquad \left(\left[\gamma \boldsymbol{P}_\pi^\intercal\right]^0 = \boldsymbol{I}\right)$$

note

$$\lim_{t\to\infty} \boldsymbol{v}_t = \lim_{t\to\infty} \left[\sum_{\tau=0}^{t-1} \left[\gamma \boldsymbol{P}_\pi^\intercal\right]^\tau \boldsymbol{r}_\pi + \left[\gamma \boldsymbol{P}_\pi^\intercal\right]^t \boldsymbol{v}_0 \right]$$

$$= \lim_{t\to\infty} \sum_{\tau=0}^{t-1} \left[\gamma \boldsymbol{P}_\pi^\intercal\right]^\tau \boldsymbol{r}_\pi + \lim_{t\to\infty} \left[\gamma \boldsymbol{P}_\pi^\intercal\right]^t \boldsymbol{v}_0$$

$$= \left[\lim_{t\to\infty} \sum_{\tau=0}^{t-1} \left[\gamma \boldsymbol{P}_\pi^\intercal\right]^\tau \right] \boldsymbol{r}_\pi + \boldsymbol{0} \qquad (\gamma < 1 \wedge \rho(\boldsymbol{P}_\pi^\intercal) = 1 \Rightarrow \lim_{t\to\infty} \left[\gamma \boldsymbol{P}_\pi^\intercal\right]^t = \boldsymbol{0})$$

$$= \left[\boldsymbol{I} - \gamma \boldsymbol{P}_\pi^\intercal\right]^{-1} \boldsymbol{r}_\pi \qquad \text{(Neumann series with } \rho(\gamma \boldsymbol{P}_\pi^\intercal) < 1)$$

$$= \boldsymbol{v}_\pi$$

take home messages

$\Rightarrow$ given some policy π, we can determine V_π

we can always proceed iteratively

$$\boldsymbol{v}_{t+1} = \boldsymbol{r}_\pi + \gamma \boldsymbol{P}_\pi^\intercal \boldsymbol{v}_t$$

or, if it isn't too expensive, directly

$$\boldsymbol{v}_\pi = \left[\boldsymbol{I} - \gamma \boldsymbol{P}_\pi^\intercal\right]^{-1} \boldsymbol{r}_\pi$$

disclaimer

we just saw that the process

for $t \in \mathbb{N}$
  for all $s \in \mathcal{S}$

$$V_{t+1}(s) = \sum_a R(s,a)\,\pi(a \mid s) + \gamma \sum_a \sum_{s'} V_t(s')\,p(s' \mid s,a)\,\pi(a \mid s)$$

is guaranteed to converge to a fixed point $V_\pi(s)$

next, we will reexamine this using different tools

to begin with, we note ...

value functions are points in an $|\mathcal{S}|$-dimensional vector space over $\mathbb{R}$

we may measure the distance between any two value functions U, V in terms of the L_∞ norm

$$\left\|U - V\right\|_\infty = \max_{s \in \mathcal{S}} \left|U(s) - V(s)\right|$$

in terms of the above vector notation, we may write this distance as

$$\left\|\boldsymbol{u} - \boldsymbol{v}\right\|_\infty$$

next, we note ...

we may write the Bellman equation as

$$\boldsymbol{v}_\pi = \boldsymbol{B}_\pi(\boldsymbol{v}_\pi)$$

where the operator $\boldsymbol{B}_\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ with

$$\boldsymbol{B}_\pi(\boldsymbol{x}) = \boldsymbol{r}_\pi + \gamma \boldsymbol{P}_\pi^\intercal \boldsymbol{x}$$

is called a Bellman policy operator or a Bellman expectation operator

the above equality $\boldsymbol{v}_\pi = \boldsymbol{B}_\pi(\boldsymbol{v}_\pi)$ implies that $\boldsymbol{v}_\pi$ is a *fixed point* of $\boldsymbol{B}_\pi$

next, we ask . . .

question

what happens to the distance between any two $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{|\mathcal{S}|}$
if we subject both points to the same Bellman operator ?

answer

this . . .

we have

$$\begin{aligned}
\left\| \boldsymbol{B}_\pi(\boldsymbol{u}) - \boldsymbol{B}_\pi(\boldsymbol{v}) \right\|_\infty &= \left\| \boldsymbol{r}_\pi + \gamma \boldsymbol{P}_\pi^\intercal \boldsymbol{u} - \boldsymbol{r}_\pi - \gamma \boldsymbol{P}_\pi^\intercal \boldsymbol{v} \right\|_\infty = \left\| \gamma \boldsymbol{P}_\pi^\intercal \left[\boldsymbol{u} - \boldsymbol{v} \right] \right\|_\infty \\
&\leqslant \left\| \gamma \boldsymbol{P}_\pi^\intercal \mathbf{1} \left\| \boldsymbol{u} - \boldsymbol{v} \right\|_\infty \right\|_\infty && \left(\boldsymbol{P}_\pi^\intercal \text{ row stochastic} \Rightarrow \boldsymbol{P}_\pi^\intercal \mathbf{1} = \mathbf{1} \right) \\
&= \gamma \left\| \boldsymbol{u} - \boldsymbol{v} \right\|_\infty && \left(0 \leqslant \gamma < 1 \right) \\
&\leqslant \left\| \boldsymbol{u} - \boldsymbol{v} \right\|_\infty
\end{aligned}$$

$\Rightarrow$ applying $\boldsymbol{B}_\pi$ brings points $\boldsymbol{u}$ and $\boldsymbol{v}$ closer together $\Leftrightarrow$ $\boldsymbol{B}_\pi$ **is a contraction operator**

given a complete metric space $(\mathcal{V}, d)$, an operator

$$f : \mathcal{V} \to \mathcal{V}$$

is a contraction mapping / contraction / contractor, if

$$\exists\, \gamma \in [0, 1) : \forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathcal{V} : d\big(f(\boldsymbol{x}), f(\boldsymbol{y})\big) \leqslant \gamma\, d(\boldsymbol{x}, \boldsymbol{y})$$

the Banach fixed point theorem establishes that

every contraction has a *unique* fixed point

$$\boldsymbol{x}_* = f(\boldsymbol{x}_*)$$

the sequence $\boldsymbol{x}_t = f(\boldsymbol{x}_{t-1}) = f^t(\boldsymbol{x}_0)$ will converge to this fixed point

$$\lim_{t \to \infty} \boldsymbol{x}_t = \boldsymbol{x}_*$$

for *any* $\boldsymbol{x}_0 \in \mathcal{V}$
the following useful upper bounds do hold

$$d(\boldsymbol{x}_t, \boldsymbol{x}_*) \leqslant \frac{\gamma^n}{1 - \gamma}\, d\big(\boldsymbol{x}_0, f(\boldsymbol{x}_0)\big) \qquad (\Leftrightarrow \text{“how many iterations from } \boldsymbol{x}_0 \text{ to get close to } \boldsymbol{x}_* \text{ ?”})$$

$$d(\boldsymbol{x}_t, \boldsymbol{x}_*) \leqslant \frac{\gamma}{1 - \gamma}\, d(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t) \qquad (\Leftrightarrow \text{“how close are we currently to to } \boldsymbol{x}_* \text{ ?”})$$

note

we just saw twice in a row that we can perform policy evaluation

$\Leftrightarrow$ given any policy π and an arbitrary initial value function $V_0(s)$, we can determine the policy's value function $V_\pi(s)$

synchronous policy evaluation

function *SynchronousPolicyEvaluation*(MDP, π, γ, ε, arbitrary V_{old})

while *TRUE*

for all $s \in \mathcal{S}$

$$V_{\text{new}}(s) \leftarrow \sum_a R(s,a)\,\pi(a \mid s) + \gamma \sum_a \sum_{s'} V_{\text{old}}(s')\,p(s' \mid s,a)\,\pi(a \mid s)$$

if $\max_{s \in \mathcal{S}} \left| V_{\text{new}}(s) - V_{\text{old}}(s) \right| \leqslant \varepsilon$

return V_{new}

$V_{\text{old}} \leftarrow V_{\text{new}}$

asynchronous policy evaluation

function *AsynchronousPolicyEvaluation*(*MDP*, π, γ, ε, arbitrary V)

while *TRUE*

$\delta \leftarrow 0$

for all $s \in \mathcal{S}$

$$h \leftarrow V(s)$$

$$V(s) \leftarrow \sum_a R(s,a)\,\pi(a \mid s) + \gamma \sum_a \sum_{s'} V(s')\,p(s' \mid s,a)\,\pi(a \mid s)$$

$$\delta \leftarrow \max(\delta, |V(s) - h|)$$

if $\delta \leqslant \varepsilon$

return V

remarks

an iteration over the s is commonly called a sweep through the state space

asynchronous or “in place” policy evaluation often converges *faster* than its synchronous counterpart, i.e. it requires fewer sweeps

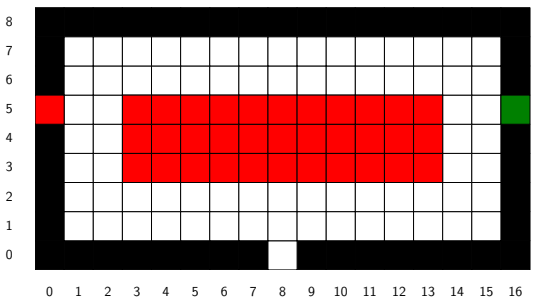
alas, the order in which states are swept over during asynchronous policy evaluation can significantly impact the rate of convergence

there are many clever ideas for how to find an optimal order but we will *not* discuss them . . . future lectures will reveal why this really would be overkill

instead, let’s look an example with synchronous updates . . .

example

a simple grid world (RL instructors love grid worlds)



states

$$\begin{aligned}\mathcal{S} &= \{s_1, s_2, s_3, \ldots, s_{108}\} \\ &= \left\{ \begin{bmatrix}0\\5\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}1\\2\end{bmatrix}, \ldots, \begin{bmatrix}16\\5\end{bmatrix} \right\}\end{aligned}$$

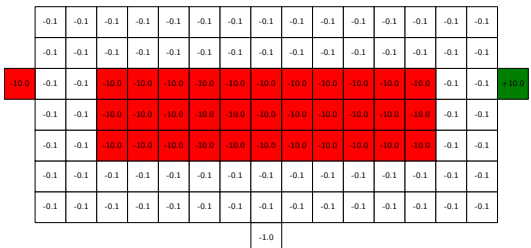
actions

$$\begin{aligned}\mathcal{A} &= \{a_1, a_2, a_3, a_4\} \\ &= \{\uparrow, \rightarrow, \downarrow, \leftarrow\} \\ &= \left\{ \begin{bmatrix}0\\+1\end{bmatrix}, \begin{bmatrix}+1\\0\end{bmatrix}, \begin{bmatrix}0\\-1\end{bmatrix}, \begin{bmatrix}-1\\0\end{bmatrix} \right\}\end{aligned}$$

final states

$$\begin{aligned}\mathcal{F}_- &= \{s_1, s_{18}, \ldots, s_{91}\} \subset \mathcal{S} \\ \mathcal{F}_+ &= \{s_{108}\} \subset \mathcal{S} \\ \mathcal{F} &= \mathcal{F}_- \cup \mathcal{F}_+\end{aligned}$$

“reward shaped” utilities $Util(s)$ of states s



reasoning

lava states $\in \mathcal{F}_-$, exit $\in \mathcal{F}_+$
entry should *not* be revisited

$\Leftrightarrow$ lava states have utility -10
the exit has a utility of $+10$
entry $[8,0]^\mathsf{T}$ has utility -1

$\Rightarrow$ utility vector

$$\boldsymbol{u} \in \mathbb{R}^{|\mathcal{S}|}$$

with

$$u_s = Util(s)$$

setting up tensor $\boldsymbol{T}$

$\boldsymbol{T} \leftarrow \mathbf{0} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}| \times |\mathcal{A}|}$

for all $a \in \mathcal{A}$
  for all $s \in \mathcal{S}$

    if $s \in \mathcal{F}$
      $T_{s,s,a} \leftarrow 1$
      continue

    $s' \leftarrow \mathit{Succ}(s, a)$

    if $s' \notin \mathcal{S}$
      $T_{s,s,a} \leftarrow 1$
    else
      $T_{s',s,a} \leftarrow 1$

$T_{:,:,a} \leftarrow \frac{T_{:,:,a}}{\sum_{s'} T_{s',:,a}}$

preparing matrix $\boldsymbol{R}$

$\boldsymbol{R} \leftarrow \mathbf{0} \in \mathbb{R}^{|\mathcal{A}| \times |\mathcal{S}|}$

for all $a \in \mathcal{A}$

$R_{a,:} = \sum_{s'} u_{s'} T_{s',:,a}$

setting up a stochastic policy matrix $\mathbf{\Pi}$

$\mathbf{\Pi} \leftarrow \mathbf{0} \in \mathbb{R}^{|\mathcal{A}| \times |\mathcal{S}|}$

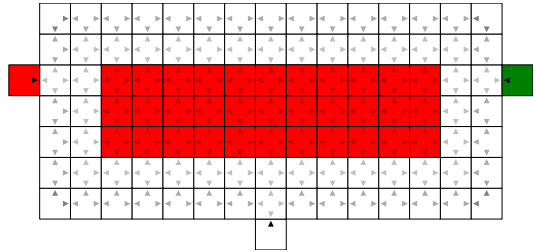
for all $a \in \mathcal{A}$
  for all $s \in \mathcal{S}$

$s' \leftarrow \mathit{Succ}(s, a)$

if $s' \in \mathcal{S}$
$\Pi_{a,s} \leftarrow 1$

$\Pi_{:,:} \leftarrow \frac{\Pi_{:,:}}{\sum_a \Pi_{a,:}}$

visualization

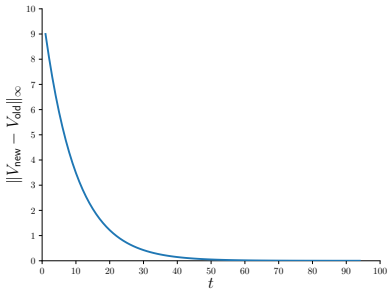


results

initial $V(s) = u(s)$, $\gamma = 0.9$, $\varepsilon = 0.0005$

-36.4	-37.7	-43.2	-46.6	-48.3	-49.0	-49.3	-49.3	-49.1	-48.5	-47.0	-43.7	-37.0	-25.2	-16.7
-42.8	-45.8	-59.4	-63.6	-65.0	-65.5	-65.7	-65.7	-65.6	-65.2	-64.1	-61.5	-53.9	-29.9	-11.7
-100.0	-60.3	-63.3	-100.0	-100.0	-100.0	-100.0	-100.0	-100.0	-100.0	-100.0	-100.0	-100.0	-41.6	+7.9	+100.0
	-50.4	-63.9	-100.0	-100.0	-100.0	-100.0	-100.0	-100.0	-100.0	-100.0	-100.0	-100.0	-51.3	-22.5	
	-43.4	-58.7	-100.0	-100.0	-100.0	-100.0	-100.0	-100.0	-100.0	-100.0	-100.0	-100.0	-52.3	-31.4	
	-35.2	-42.1	-58.1	-63.0	-64.6	-65.1	-65.0	-64.6	-65.0	-65.0	-64.5	-62.5	-56.9	-38.5	-29.4
	-31.6	-34.7	-41.7	-45.8	-47.7	-48.3	-47.8	-45.6	-47.7	-48.2	-47.4	-45.2	-40.4	-32.1	-27.8
								-41.1							

convergence behavior



results

initial $V(s) = u(s)$, $\gamma = 0.5$, $\varepsilon = 0.0005$

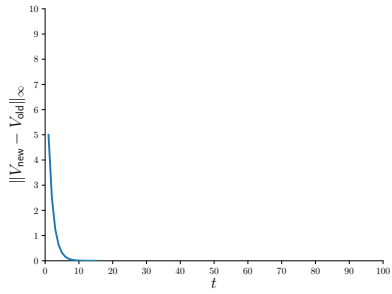
-0.7	-0.8	-1.6	-1.9	-1.9	-1.9	-1.9	-1.9	-1.9	-1.9	-1.9	-1.8	-1.6	-0.7	-0.2
-1.7	-2.1	-6.4	-7.0	-7.1	-7.1	-7.1	-7.1	-7.1	-7.1	-7.1	-7.0	-6.4	-1.6	+0.3
-20.0	-6.5	-7.0	-20.0	-20.0	-20.0	-20.0	-20.0	-20.0	-20.0	-20.0	-20.0	-20.0	-5.6	+4.2	+20.0
-2.7	-7.1	-20.0	-20.0	-20.0	-20.0	-20.0	-20.0	-20.0	-20.0	-20.0	-20.0	-20.0	-6.7	-0.7
-1.7	-6.4	-20.0	-20.0	-20.0	-20.0	-20.0	-20.0	-20.0	-20.0	-20.0	-20.0	-20.0	-6.3	-1.4
-0.8	-1.9	-6.4	-7.0	-7.1	-7.1	-7.1	-7.1	-7.1	-7.1	-7.1	-7.0	-6.4	-1.9	-0.7
-0.5	-0.8	-1.6	-1.8	-1.9	-1.9	-1.9	-1.8	-1.9	-1.9	-1.9	-1.8	-1.6	-0.8	-0.5
							-1.0							

observe

many values seem identical, but this is due to rounding

all values do indeed differ after the 4th (!) decimal place

convergence behavior



question

why does the $\gamma = 0.5$ version converge faster than the version with $\gamma = 0.9$?

answer

given what we saw above, you might want to figure this out all by yourself ;-)

question

OK, but now for real: what about $V_*(s)$ and $\pi_*(s)$?

answer

...

Bellman optimality equations

Bellman optimality equation for $V_*(s)$

$$
\begin{aligned}
V_*(s) &= \max_a Q_*(s,a) \\
&= \max_a \mathbb{E}\Big[R_1 + \gamma V_*(S_1) \;\Big|\; S_0 = s, A_0 = a\Big] \\
&= \max_a R(s,a) + \gamma \sum_{s'} V_*(s')\, p(s' \mid s,a)
\end{aligned}
$$

Bellman optimality equation for $Q_*(s,a)$

$$Q_*(s,a) = \mathbb{E}\Big[R_1 + \gamma \max_{a'} Q_*(S_1,a') \;\Big|\; S_0 = s, A_0 = a\Big]$$

$$= R(s,a) + \gamma \sum_{s'} \max_{a'} Q_*(s',a')\, p(s' \mid s,a)$$

note

because of the occurrence of the “max” operator neither system of equations is linear anymore

if we could determine $Q_*(s, a)$, we could easily determine $V_*(s)$

if we knew $Q_*(s, a)$ or $V_*(s)$, we could easily get $\pi_*(s)$ because ...

optimal policy $\pi_*(s)$

$$\pi_*(s) = \underset{a}{\text{argmax}}\ Q_*(s,a)$$

$$= \underset{a}{\text{argmax}}\ R(s,a) + \gamma \sum_{s'} V_*(s')\, p(s' \mid s,a)$$

question

so how can we proceed ?

answer

will be given next time . . .

summary

we now know about

aspects of working with deterministic policies

Bellman's principle of optimality

(iterative) policy evaluation

Bellman optimality equations