Reinforcement Learning

Lecture 3

Planning

Closed Form Solution of Value Function

Iterative Policy Evaluation

Robert Peharz

Institute of Theoretical Computer Science
Graz University of Technology
Winter Term 2023/24

Recap

A **Markov Decision Process (MDP)** is a 3-tuple

$$\boxed{(\mathcal{S}, \mathcal{A}, p)}$$

where

$\mathcal{S}$ is a **state space** (**set of states** of environment)

$\mathcal{A}$ is an **action space** (**set of actions** of agent)

p is called the **dynamics** of the MDP. Formally, it is a **conditional probability distribution**

$$p(s', r \mid s, a)$$

where

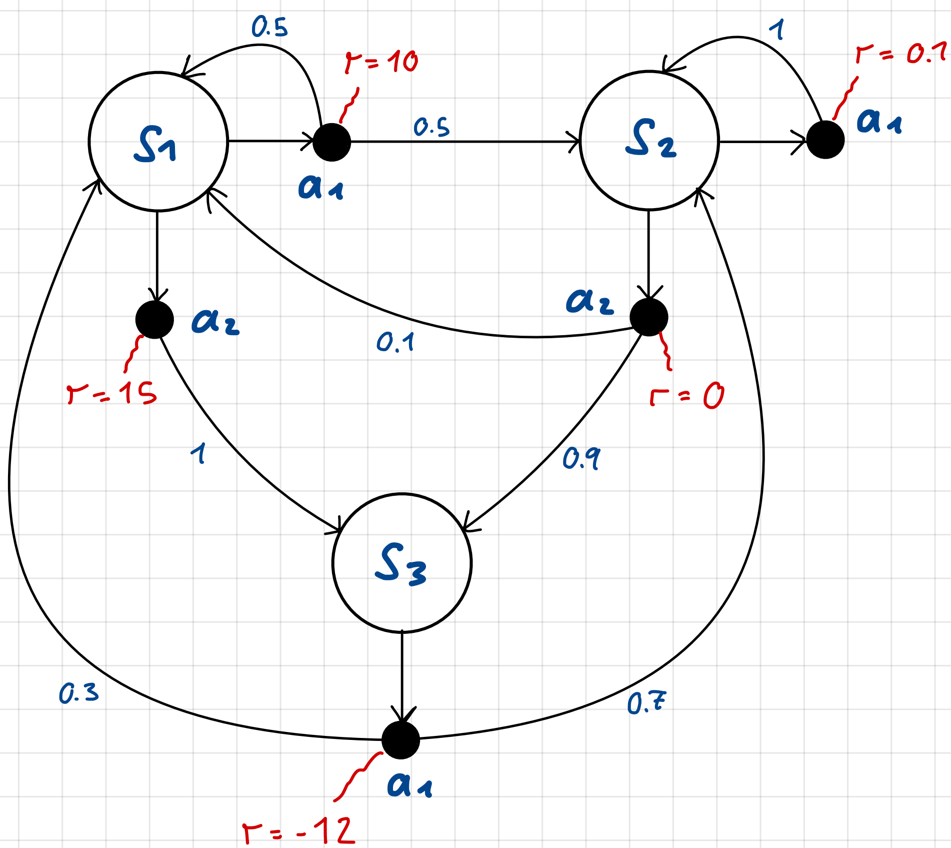
$s' \in \mathcal{S}$ is the value of the **next state** S_{t+1},

$r \in \mathbb{R}$ is the value of the **reward** R_{t+1},

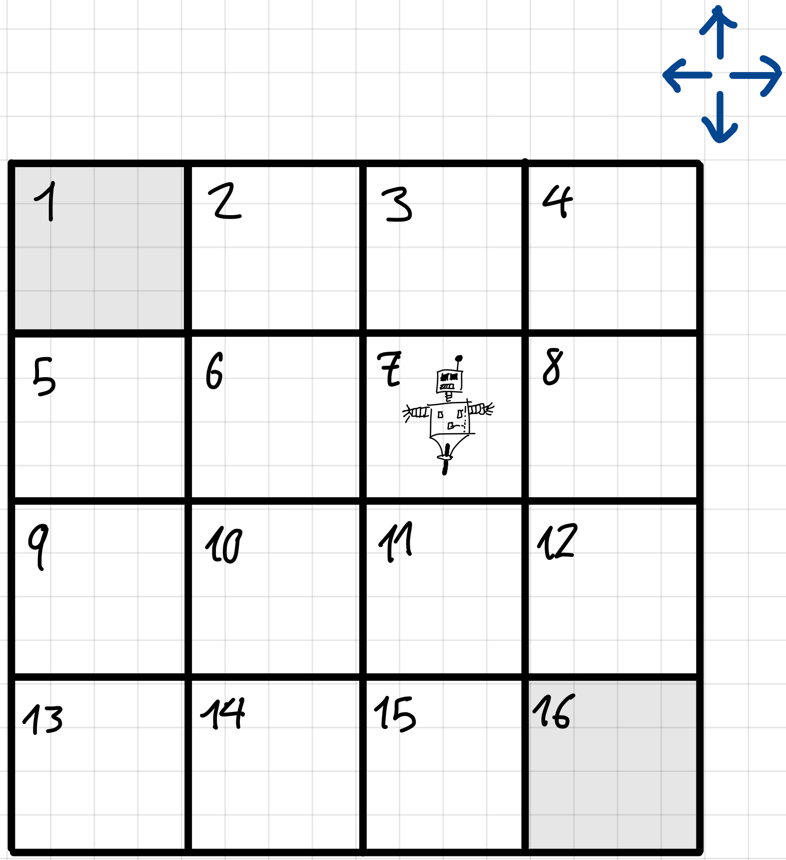
$s \in \mathcal{S}$ is the value of the **current state** S_t,

$a \in \mathcal{A}$ is the value of the **selected action** A_t.

Recap



Transition Diagram



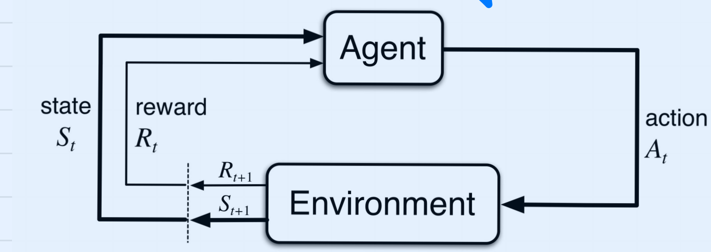
Grid World

Recap

Policy (agent's behavior)

$$\boxed{\pi(a \mid s)}$$

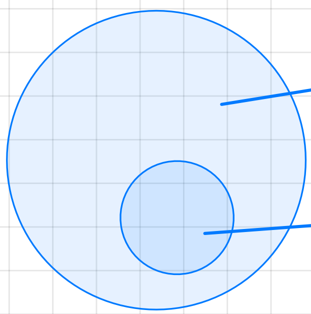
policy



MDP dynamics

(conditional distribution over actions given current state)

Subsumes deterministic policies



stochastic policies

deterministic policies

$$\pi_{sto}(a \mid s) := \begin{cases} 1 & \text{if } \pi_{det}(s) = a \\ 0 & \text{otherwise} \end{cases}$$

Recap

The **discounted return** at time t is defined as

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

$0 \leq \gamma \leq 1$

The **value function** $v_\pi : \mathcal{S} \to \mathbb{R}$ is defined as

$$v_\pi(s) := \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right]$$

"The expected discounted return when following policy π, and starting from $s \in \mathcal{S}$ (t is any arbitrary time step)."

Bellman Expectation Equation

$$v_\pi(s) = \mathbb{E}_{A_t, R_{t+1}, S_{t+1}}\left[R_{t+1} + \gamma\, v_\pi(S_{t+1}) \mid S_t = s\right]$$

Image: wikipedia.org

Recap

$$v_\pi(s) = \mathbb{E}_{A_t, R_{t+1}, S_{t+1}}\left[R_{t+1} + \gamma\, v_\pi(S_{t+1}) \mid S_t = s\right]$$

$$v_\pi(s) = \sum_a \pi(a|s)\left[\sum_{s',r} p(s',r|s,a)\left(r + \gamma\, v_\pi(s')\right)\right]$$

$$\vdots$$

$$= r(s) + \gamma \sum_{s'} p(s'|s)\, v_\pi(s')$$

where

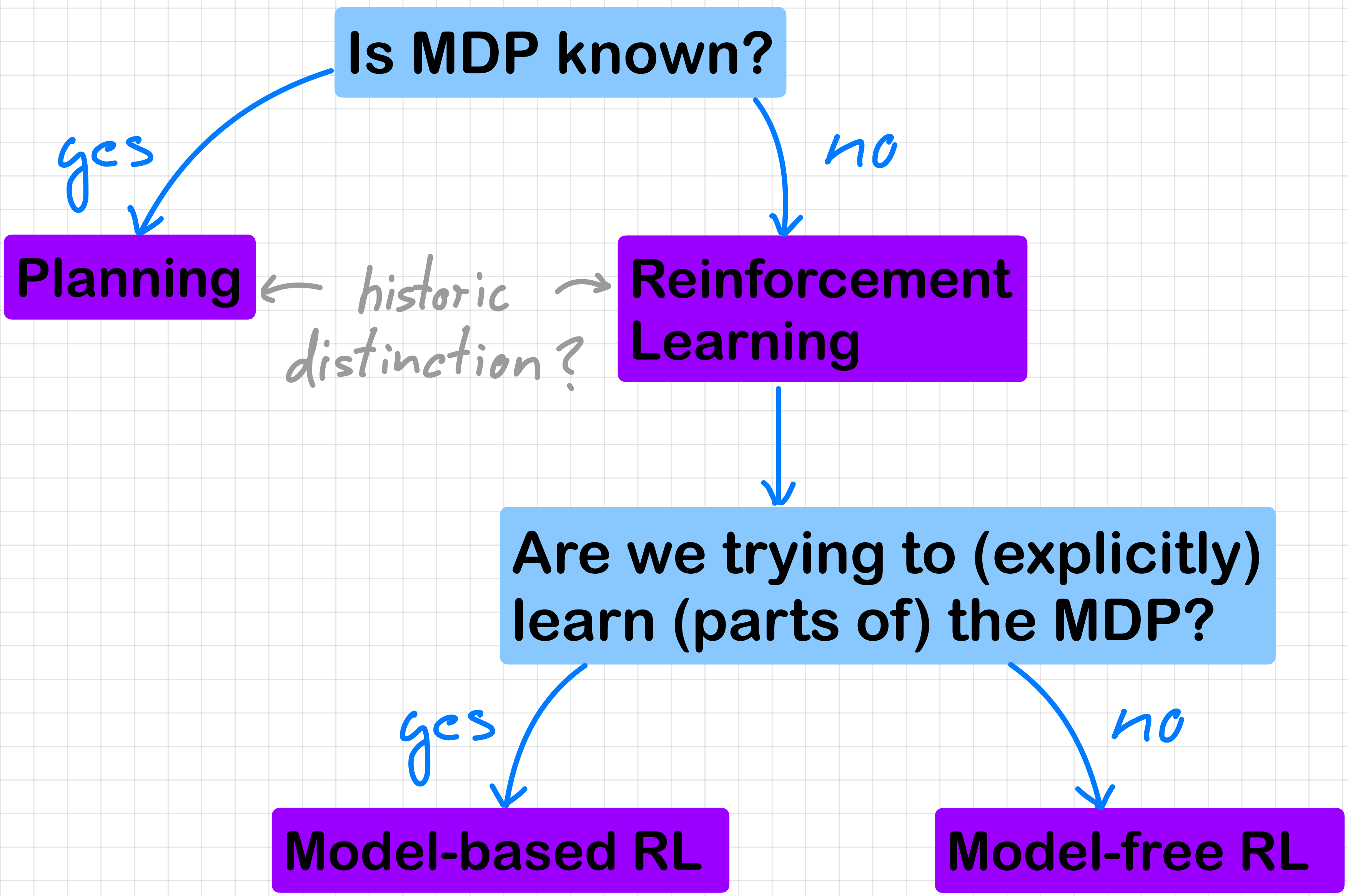
$$\overbrace{r(s)}^{\text{starting at } s} = \sum_a \pi(a|s) \underbrace{\sum_{s',r} p(s',r|s,a)\, r}_{} = \sum_a \pi(a|s)\, \overbrace{r(s,a)}^{\text{starting at } s \text{ with } a}$$

expected reward

combined state transition

$$p(s'|s) = \sum_a \sum_r \overbrace{\underbrace{\pi(a|s)}_{\text{policy}}\ \underbrace{p(s',r|s,a)}_{\text{dynamics}}}^{p(s',a,r|s)\ \text{(chain rule)}}$$

Planning, Model-based RL, Model-free RL



Fundamental Problems in Planning & RL

Policy Evaluation, Prediction – How good is my policy?

given a policy π, compute the value function v_π

Control – What is the best policy?

How to improve a given policy π, or even find an optimal policy?

The two problems are related, and the algorithms to solve them are similar

Planning: Policy Evaluation, Prediction

Solving the Bellman Equation(s)

- note that we actually have one Bellman equation per state:

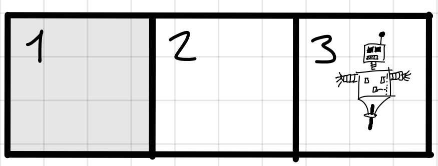
$$v_\pi(s_1) = r(s_1) + \gamma \sum_{s'} p(s'|s_1)\, v_\pi(s')$$

$$v_\pi(s_2) = r(s_2) + \gamma \sum_{s'} p(s'|s_2)\, v_\pi(s')$$

$$\vdots$$

$$v_\pi(s_N) = r(s_N) + \gamma \sum_{s'} p(s'|s_N)\, v_\pi(s') \qquad \text{(assuming } N \text{ states)}$$

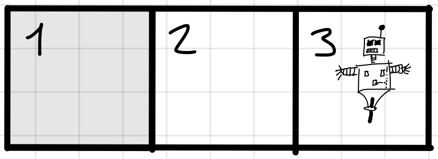
- assume a very simple grid world



  - states $\mathcal{S} = \{1, 2, 3\}$, state 1 is terminal
  - actions $\mathcal{A} = \{\leftarrow, \rightarrow\}$, move agent (deterministic)
  - reward -1 non-terminal states, 0 for terminal

- assume a random walk policy $\pi(a|s) = 0.5$ for all s, a

Solving the Bellman Equation(s) cont'd



- states $\mathcal{S} = \{1,2,3\}$, state 1 is terminal
- actions $\mathcal{A} = \{\leftarrow, \rightarrow\}$, move agent (deterministic)
- reward -1 non-terminal states, 0 for terminal

combined state transition as matrix

$|\mathcal{S}| \times |\mathcal{S}|$-matrix

$$P = \begin{pmatrix} p(1|1) & p(2|1) & p(3|1) \\ p(1|2) & p(2|2) & p(3|2) \\ p(1|3) & p(2|3) & p(3|3) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0.5 & 0.5 \end{pmatrix}$$

stochastic matrix

(non-negative, rows sum to 1)

e.g. $$p(1|2) = \pi(\leftarrow|2)\, p(1|2,\leftarrow) + \pi(\rightarrow|2)\, p(1|2,\rightarrow)$$
$$= 0.5 \times 1 + 0.5 \times 0 = 0.5$$

expected immediate reward as vector

$$r = \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix}$$

$|\mathcal{S}|$-dim vector

Solving the Bellman Equation(s) cont'd

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0.5 & 0.5 \end{pmatrix} \qquad r = \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix}$$

Bellman equation for state 1

$$v_\pi(1) = r(1) + \gamma\left[p(1|1)\, v_\pi(1) + p(2|1)\, v_\pi(2) + p(3|1)\, v_\pi(3)\right]$$

$$= 0 + \gamma\left[1 \quad v_\pi(1) + 0 \quad v_\pi(2) + 0 \quad v_\pi(3)\right]$$

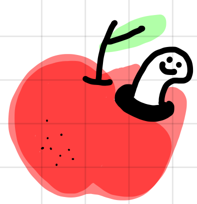
$$= 0 + \gamma\, v_\pi(1)$$

$$(1-\gamma)\, v_\pi(1) = 0$$

Thus, if $\gamma < 1$ (discounting) $\underline{v_\pi(1) = 0}$

For $\gamma = 1$, Bellman equation is vacuous here...

Bellman equation is <u>not</u> the definition of v_π !



Even for $\gamma = 1$ we can argue that $v_\pi(1) = 0$, since

$v_\pi(1) = \mathbb{E}_\pi[G_t \mid s = 1] = 0$ as 1 is a terminal state

Solving the Bellman Equation(s) cont'd

Bellman equations for states 2 and 3

$$v_\pi(2) = r(2) + \gamma \left[\cancel{p(1|2)\, v_\pi(1)} + \cancel{p(2|2)\, v_\pi(2)} + p(3|2)\, v_\pi(3) \right]$$

$$v_\pi(3) = r(3) + \gamma \left[\underbrace{\cancel{p(1|3)\, v_\pi(1)}}_{=0} + p(2|3)\, v_\pi(2) + p(3|3)\, v_\pi(3) \right]$$

$$\left.\begin{aligned} v_\pi(2) &= -1 + \gamma\, 0.5\, v_\pi(3) \\ v_\pi(3) &= -1 + \gamma\, 0.5\, v_\pi(2) + \gamma\, 0.5\, v_\pi(3) \end{aligned}\right\} \begin{array}{l} \text{2 equations} \\ \text{2 unknowns} \end{array}$$

Assume $\gamma = 1$

$$v_\pi(3) = -1 + 0.5\, \overbrace{\left(-1 + 0.5\, v_\pi(3)\right)}^{v_\pi(2)} + 0.5\, v_\pi(3)$$

$$= -1.5 + 0.75\, v_\pi(3)$$

$$v_\pi(3) = \frac{-1.5}{0.25} = -6$$

$$v_\pi(2) = -1 + 0.5\, v_\pi(3) = -4$$

Analytic Solution of Bellman Equation

$$p(s'|s) = \sum_{r,a} p(s',r|s,a)\,\pi(a|s)$$

$$v_\pi(s) = r(s) + \gamma \sum_{s'} p(s'|s)\, v_\pi(s')$$

Assume an arbitrary enumeration of states $\mathcal{S} = \{s_1, s_2, \dots s_N\}$

$$\underline{v} = \begin{pmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ \vdots \\ v_\pi(s_N) \end{pmatrix} \quad \underline{r} = \begin{pmatrix} r(s_1) \\ r(s_2) \\ \vdots \\ r(s_N) \end{pmatrix} \quad P = \begin{pmatrix} p(s_1|s_1) & \dots & p(s_N|s_1) \\ \vdots & \ddots & \vdots \\ p(s_1|s_N) & \dots & p(s_N|s_N) \end{pmatrix}$$

We can write the Bellman equations as $\underline{v} = \underline{r} + \gamma P \underline{v}$

Thus,

$$\underline{v} - \gamma P \underline{v} = \underline{r}$$

$$(I - \gamma P)\,\underline{v} = \underline{r}$$

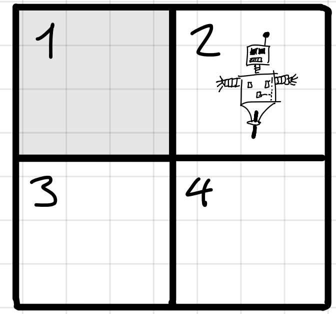
$$\underline{v}_\pi = (I - \gamma P)^{-1}\, \underline{r}$$

Example: 4 State Grid World

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.25 & 0.5 & 0 & 0.25 \\ 0.25 & 0 & 0.5 & 0.25 \\ 0 & 0.25 & 0.25 & 0.5 \end{pmatrix} \qquad \underline{r} = \begin{pmatrix} 0 \\ -1 \\ -1 \\ -1 \end{pmatrix}$$

$$\pi(a|s) = 0.25$$

$$\gamma = 0.9999$$



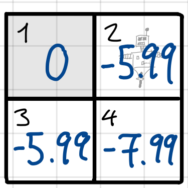

```
>>> import numpy as np
>>> P = np.array([[1., 0., 0., 0.], [0.25, 0.5, 0., 0.25], [0.25, 0., 0.5, 0.25], [0., 0.25, 0.25, 0.5]])
>>> P
array([[1.  , 0.  , 0.  , 0.  ],
       [0.25, 0.5 , 0.  , 0.25],
       [0.25, 0.  , 0.5 , 0.25],
       [0.  , 0.25, 0.25, 0.5 ]])
>>> r = np.array([0., -1., -1., -1.])
>>> r
array([ 0., -1., -1., -1.])
>>> gamma = 0.9999
>>> np.linalg.inv(np.eye(4) - gamma*P) @ r
array([-2.25605909e-16, -5.99660198e+00, -5.99660198e+00, -7.99520280e+00])
>>>
```

$$\underline{v} = (I - \gamma P)^{-1} \underline{r} \approx \begin{pmatrix} 0 \\ -5.99 \\ -5.99 \\ -7.99 \end{pmatrix} \longleftarrow$$

<u>**double check**</u>

$\underline{v}$ must satisfy Bellman equation

$$\underline{v} = \underline{r} + \gamma P \underline{v}$$



Uniqueness of Solution

Bellman Equation (three different notations)

$$v_\pi(s) = \mathbb{E}_{A_t, R_{t+1}, S_{t+1}}\left[R_{t+1} + \gamma\, v_\pi(S_{t+1}) \mid S_t = s \right]$$

$$v_\pi(s) = r(s) + \sum_{s'} p(s'|s)\, \gamma\, v_\pi(s')$$

$$\underline{v} = \underline{r} + \gamma P \underline{v}$$

Is there always a (unique) solution?

- For $\gamma = 1$, $(I - \gamma P)$ is singular and the solution is not unique
- For $\gamma < 1$, $(I - \gamma P)$ is always invertible — solution unique
- Thus, discounting makes the problem well-posed

Planning: Iterative Policy Evaluation

Iterative Policy Evaluation

- solving the Bellman equation via matrix inversion does not scale to large MDPs (cubic complexity)
- also not applicable to later problems, e.g. control
- iterative method ?
- reconsider the Bellman equation:

$$v_\pi(s) = r(s) + \gamma \sum_{s'} p(s'|s)\, v_\pi(s')$$

- this looks like an **update rule!**

Iterative Policy Evaluation

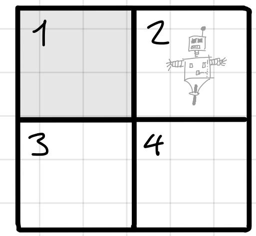
- initialize $v(s)$ arbitrarily (e.g. all 0)
- repeat
  - for $s \in \mathcal{S}$ do $v_{new}(s) \leftarrow r(s) + \gamma \sum_{s'} p(s'|s)\, v(s')$
  - if $\forall s: v_{new}(s) \approx v(s) \longrightarrow$ break
  - $v \leftarrow v_{new}$

Iterative Policy Evaluation

$\gamma = 0.9999$

iteration	$\underline{V}$: 1	2	3	4
0	0	0	0	0
1	0	-1	-1	-1
2	0	-2.4	-2.4	-2.9
⋮	⋮			
10	0	-4.8	-4.8	-6.3
⋮	⋮			
20	0	-5.8	-5.8	-7.6
⋮	⋮			
50	0	-5.99	-5.99	-7.99

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.25 & 0.5 & 0 & 0.25 \\ 0.25 & 0 & 0.5 & 0.25 \\ 0 & 0.25 & 0.25 & 0.5 \end{pmatrix} \qquad \underline{r} = \begin{pmatrix} 0 \\ -1 \\ -1 \\ -1 \end{pmatrix}$$



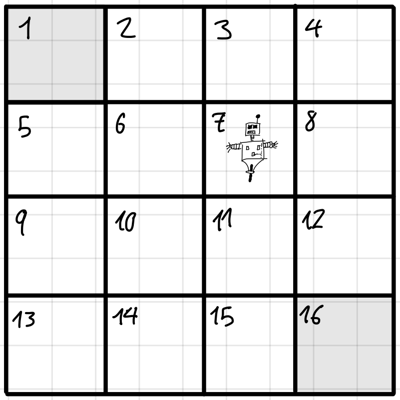
$$\underline{V} = \underline{r} + \gamma P \underline{V}$$

Recall:

$$\underline{V} = (I - \gamma P)^{-1} \underline{r} \approx \begin{pmatrix} 0 \\ -5.99 \\ -5.99 \\ -7.99 \end{pmatrix}$$

Does iterative policy evaluation **always** converge to V_π?

Iterative Policy Evaluation, Larger Gridworld



- actions ↕↔ move agent deterministically
- reward −1 whenever on non-terminal (white) cell
- discount factor $\gamma = 0.999$
- policy: $\pi(a|s) = 0.25$ (random walk)

After K iterations:

$k=0$

0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

$k=1$

0	-1	-1	-1
-1	-1	-1	-1
-1	-1	-1	-1
-1	-1	-1	0

$k=2$

0	-1.7	-2	-2
-1.7	-2	-2	-2
-2	-2	-2	-1.7
-2	-2	-1.7	0

...

$K=10$

0	-6.1	-8.3	-8.9
-6.1	-7.7	-8.4	-8.3
-8.3	-8.4	-7.7	-6.1
-8.9	-8.3	-6.1	0

...

$K=199$

0	-13.8	-19.6	-21.6
-13.8	-17.7	-19.6	-19.6
-19.6	-19.6	-17.7	-13.8
-21.6	-19.6	-13.8	0

$K=200$

0	-13.8	-19.6	-21.6
-13.8	-17.7	-19.6	-19.6
-19.6	-19.6	-17.7	-13.8
-21.6	-19.6	-13.8	0

$\approx V_\pi$

Some Mathematical Tools

Lipschitz function

Let V be any vector space and $T: V \to V$ be a function from V to V. If there is a constant L such that

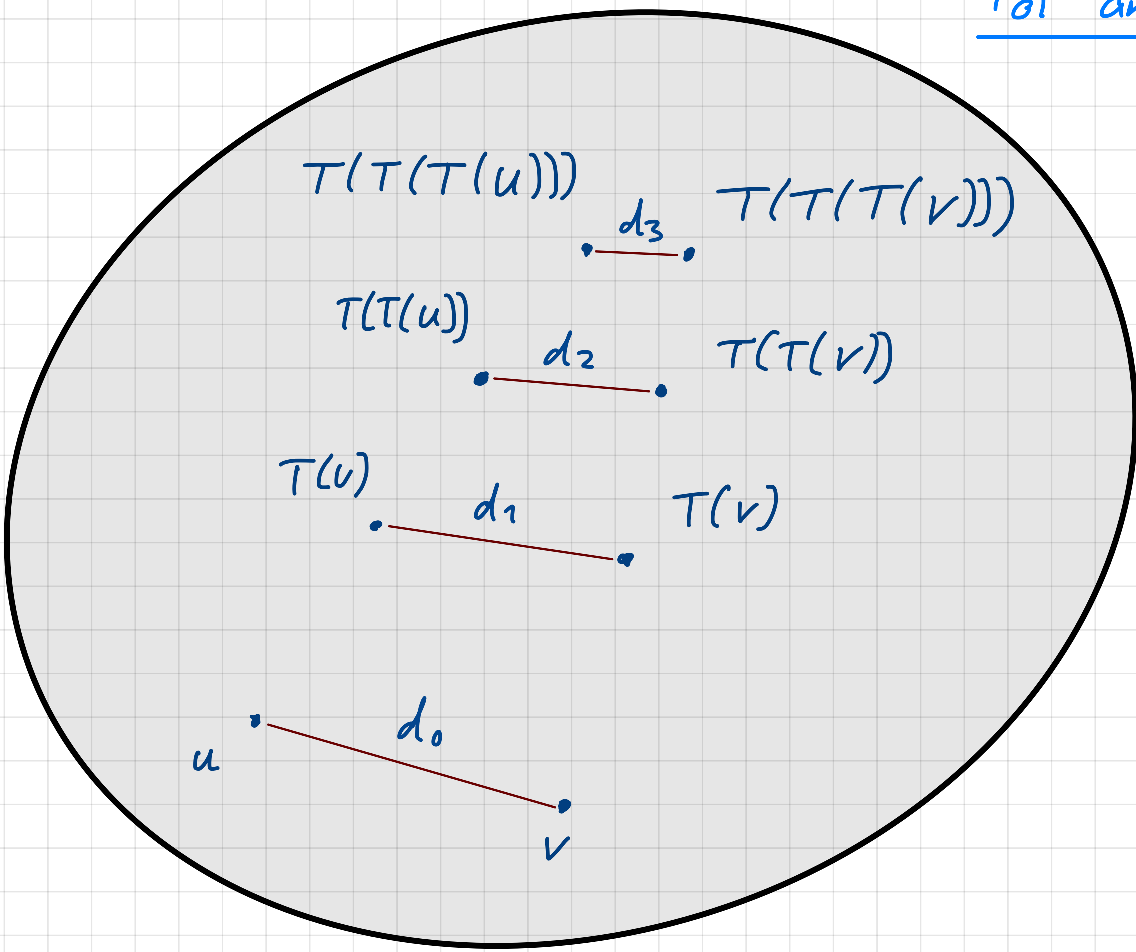
$$\|T(v) - T(v')\| \leq L \|v - v'\| \qquad \forall\, v, v' \in V$$

then T is called L-**Lipschitz**, where $\|\cdot\|$ is any norm.

Contraction (Contraction mapping)

Let V be any vector space. An L-Lipschitz function $T: V \to V$ with $L < 1$ is called a **contraction** (L-contraction).

Contraction



For any two distinct points u, v:

$$d_0 = \|u - v\|$$

$$d_1 = \|T(u) - T(v)\|$$

$$d_2 = \|T(T(u)) - T(T(v))\|$$

$$\vdots$$

$$d_0 > d_1 > d_2 > d_3 > \dots$$

Banach's Fixed Point Theorem

Fixed point

Let V be some vector space and let $T: V \to V$ be a function. A vector $v \in V$ is called a **fixed point** of T if $v = T(v)$, i.e. if T maps v onto itself.

Let V be some vector space and let $T: V \to V$ be a contraction. Then T has a unique fixed point $v^* \in V$. Moreover, for any $v_0 \in V$, the sequence defined via

$$v_n = T(v_{n-1})$$

converges to v^*, i.e. $v_n \to v^*$ as $n \to \infty$.

Details

Norm: We tacitly assumed some norm $\|\cdot\|$ of vector space V.
A common choice is the **max norm** (uniform norm, infinity norm)

$$\|\underline{v}\|_\infty = \max_i |v_i|$$

Complete space: The fixed-point theorem requires V to be **complete**, i.e. that every Cauchy sequence converges to some element in V.
For example, the real numbers $\mathbb{R}$ are complete.
The rational numbers $\mathbb{Q}$ are not complete; e.g., there are sequences in $\mathbb{Q}$ which converge to a irrational number (e.g. π).

A complete vector space with a norm is a **Banach space**.
Thus, the Banach fixed-point theorem says that iterating a contraction T in a Banach space always converges to a unique fixed-point.

Iterative Policy Evaluation

For $\gamma < 1$, Iterative Policy Evaluation always converges to v_π, i.e. for any MDP, any π and any initialisation of v_π.

- for $\gamma < 1$, the Bellman Equation is a contraction. (see HW1 in KU)
- thus, iterating it will converge to a unique fixed-point
- v_π is a fixed-point, thus Iterative Policy Evaluation converges to v_π

Contraction arguments like this are frequently used in RL theory!

Summary — Planning: Policy Evaluation, Prediction

- We have learned about two methods for evaluating a policy π
  - closed form

$$\underline{v}_\pi = (I - \gamma P)^{-1} \underline{r}$$

  - iterative policy evaluation

init $\underline{v}_0$

iterate $\underline{v}_n = \underline{r} + \gamma P \underline{v}_{n-1}$

- **Policy evaluation** just tells us how good π is, for each s.
- But how to actually **learn** π?
- Or, given some π, how to **improve** it? (next lecture)