Markov Decision Processes

Paolo Turrini

Department of Computing, Imperial College London

Introduction to Artificial Intelligence

The lectures

- The agent and the world (**Knowledge Representation**)
  - Actions and knowledge
  - Inference
- Good decisions (**Risk and Decisions**)
  - Chance
  - Gains
- Good decisions in time (**Markov Decision Processes**)
  - Chance and gains in time
  - Patience
  - Finding the best strategy
- Learning from experience (**Reinforcement Learning**)
  - Finding a reasonable strategy

Outline

- Adjusting the pace
- Plans, again
- Policies

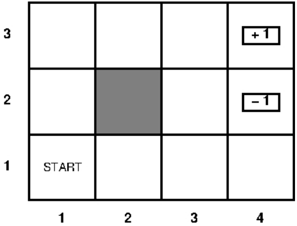
Markov Decision Processes (ii)

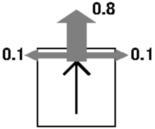
Keep making the right decisions
and results will come (almost certainly)

The book

Stuart Russell and Peter Norvig
Artificial Intelligence: a modern approach
Chapter 17

Recall...





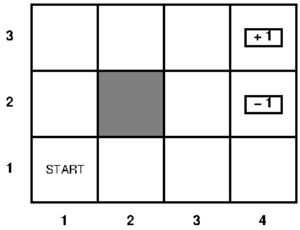
A Markov Decision Process is a sequential decision problem with:

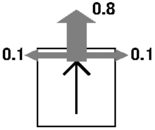
- a fully observable environment
- stochastic actions
- a Markovian transition model
- discounted rewards

In words...

- **Fully observable environment** means that I know which state I am in;
- **Stochastic actions** means that I can only choose an intended direction, but the consequences of my actions are determined probabilistically;
- **Markovian transition model** means that the future is only determined by my action now and the state I am in now: the past does not matter;
- **Discounted rewards** means that I value the rewards I receive depending on how far away they are.

Plans



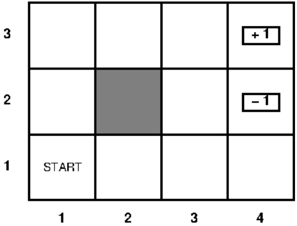


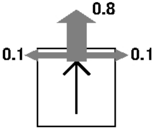
A plan is a finite sequence of intended moves, from the starting state.

So, if the action set is $\{Up, Down, Left, Right\}$,
then $[Up]$ is a plan, $[Right, Right, Left, Left]$ is a plan etc.

Recall: The effect of an action is a probability, the effect of a plan is the product of the effect of its actions.

Makings plans





- The probability that $[Up, Up, Right, Right, Right]$ gets us to $+1$ is NOT 0.8^5.
- There is a small chance of $[Up, Up, Right, Right, Right]$ accidentally reaching $+1$ by going the other way round.
- So the probability is actually $0.8^5 + (0.1^4 \times 0.8)$

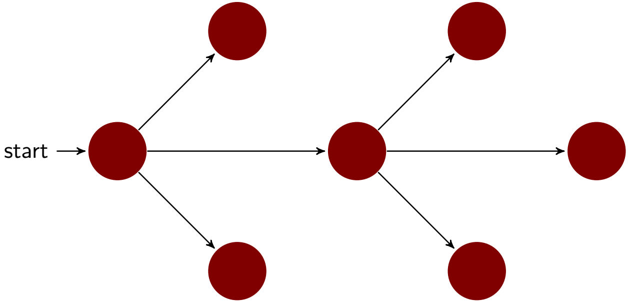
The value of a plan

The value of a plan p, from state s is the expected utility of the resulting sequences, appropriately discounted.

$$v^p(s) = E[\sum_{t=0}^{\infty} \gamma^t r(S_t)]$$

- Calculate the utility of the sequences you can actually perform, with the appropriately discounted rewards, times the probability of reaching them
- Add these numbers
- Forget about the rest

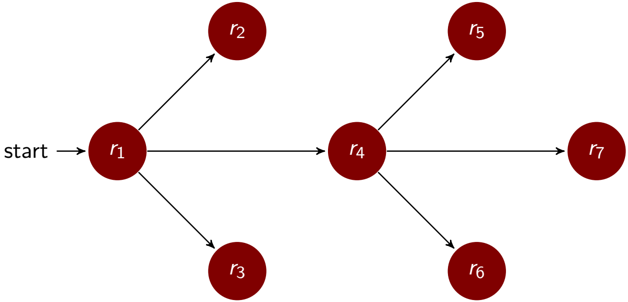
An MDP



States and transitions.

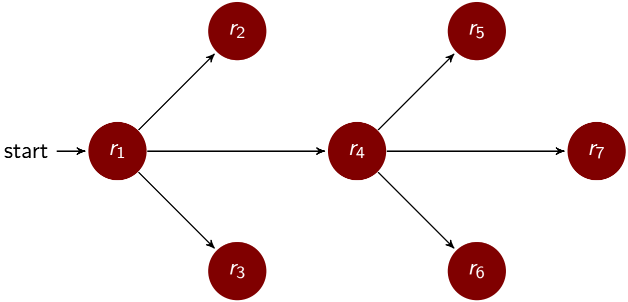
Remember: states are not necessarily the same as squares.

Time, risk and patience



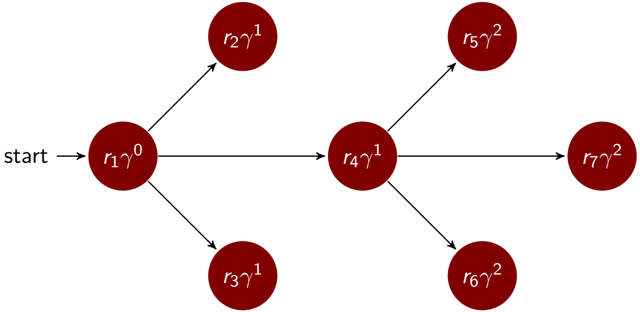
Rewards are what you get by visiting states. They can be any number, but usually they are small, negative and uniform at non-terminal states.

Time, risk and patience



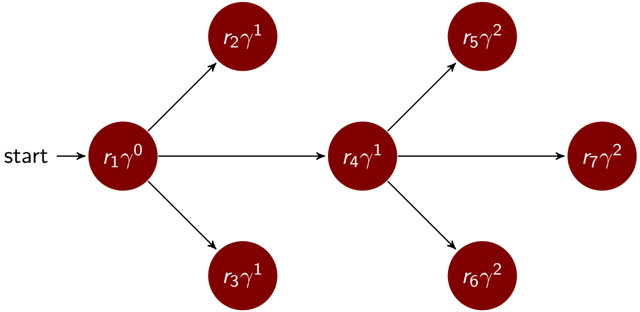
These are objective rewards: what you are going to get at the state. They are not subjective rewards: what you think these rewards are actually worth **from where you are**.

Time, risk and patience



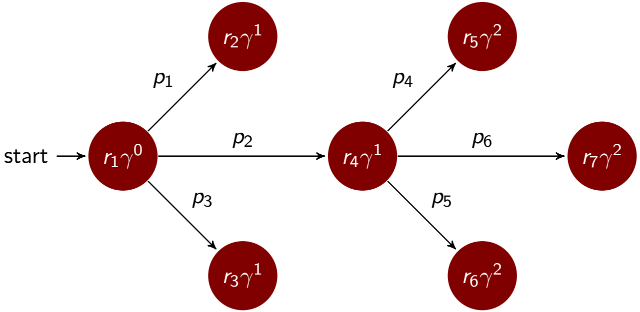
You have a multiplicative discounting $\gamma \in [0,1]$, according to which you weigh rewards.
The idea is that you prefer five today to five tomorrow.

Time, risk and patience



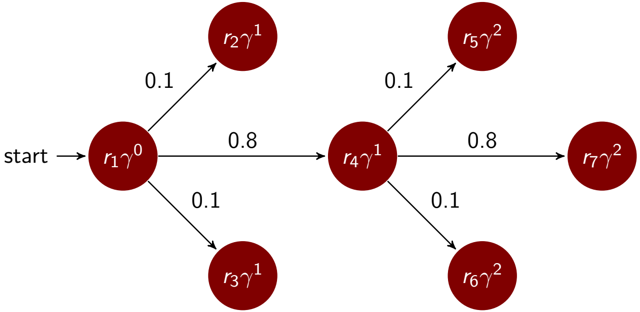
You multiply a reward by γ^n if it takes you n steps to reach it.
This is your perceived reward, and it's what really matters.

Time, risk and patience



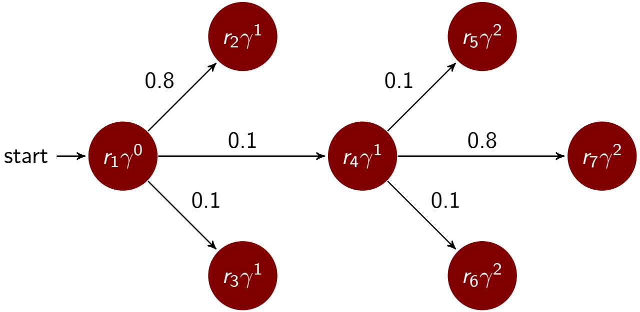
You can make plans. But being a stochastic agent each action has a probabilistic effect.

Time, risk and patience



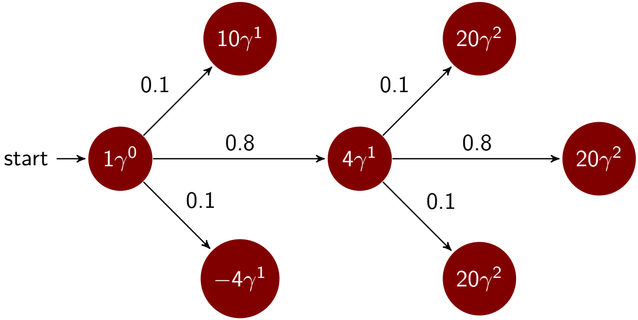
This is $[Right, Right]$, with 0.8 on the intended direction and 0.1 otherwise.

Time, risk and patience



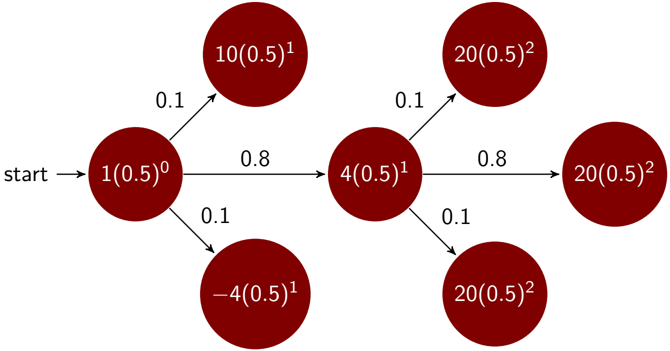
$[Up, Right]$

Time, risk and patience



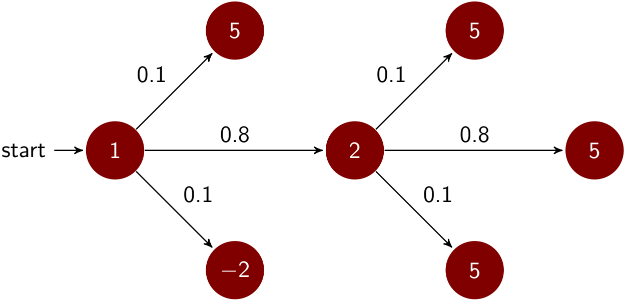
Let's throw in some arbitrary numbers corresponding to rewards.

Time, risk and patience



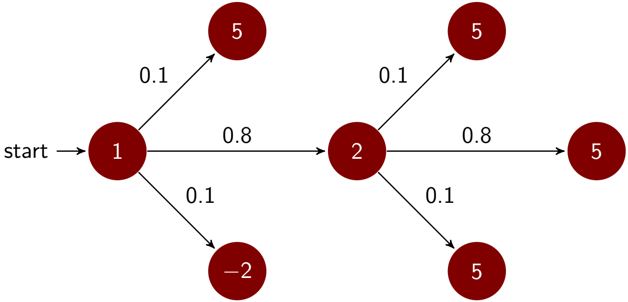
Let us assume $\gamma = 0.5$. Again, it's an arbitrary choice.

Time, risk and patience



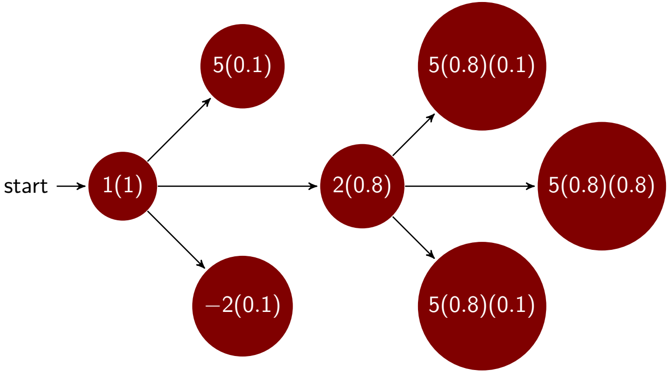
These are the perceived rewards from the starting square.

Time, risk and patience



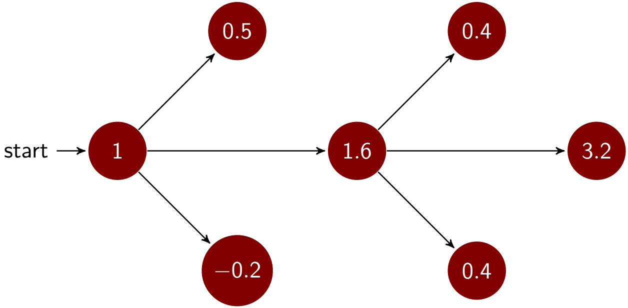
Now we are interested in seeing how likely these rewards are.
The **expected perceived rewards**.

Time, risk and patience



The probabilities are a sort of extra discount...

Time, risk and patience

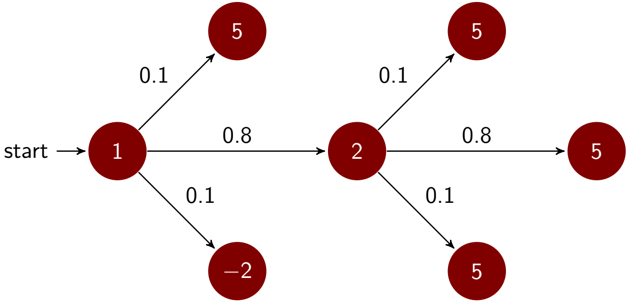


So this is the collection of perceived rewards I'm expected to get. Putting all together is: 6.9. This is the value of $[Right, Right]$, in this MDP, with $\gamma = 0.5$

Some plans are better than others

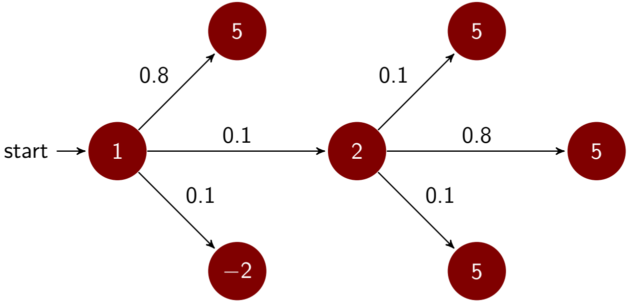
- Estimating the expected utility of two different plans is going to tell us which one NOT to choose.
- Ideally we want to find the best plan, but we will have often have to take a decision and going with the most satisfactory one we have found.

Time, risk and patience



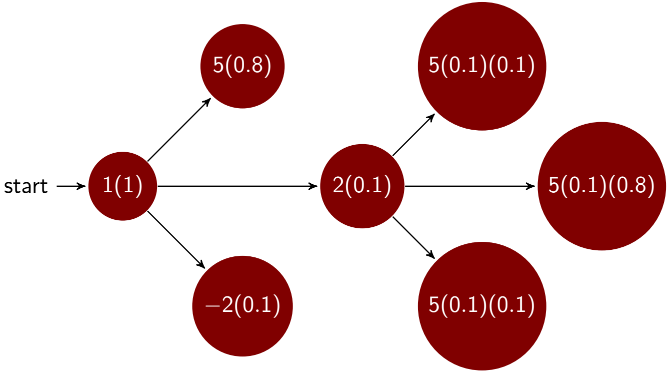
There is only so much we can do here.
Because of the rewards, the second move is irrelevant:
[a, b] is the same as [a, c].
Which means we only need to check two more moves.

Time, risk and patience



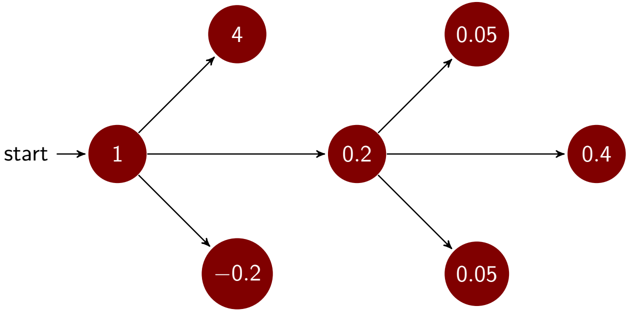
Let's go Up first.

Time, risk and patience



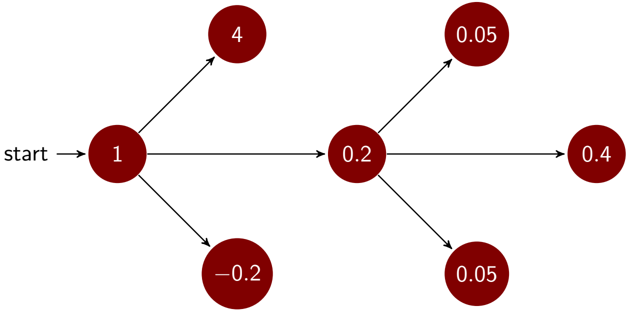
Including probabilities...

Time, risk and patience



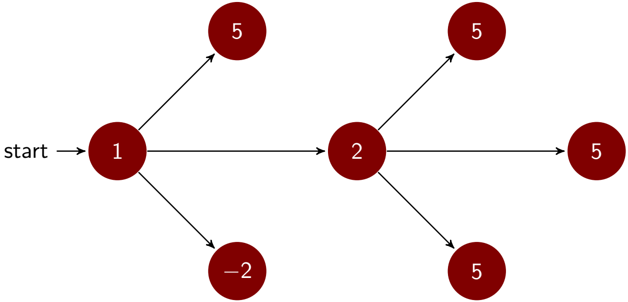
Including probabilities...

Time, risk and patience



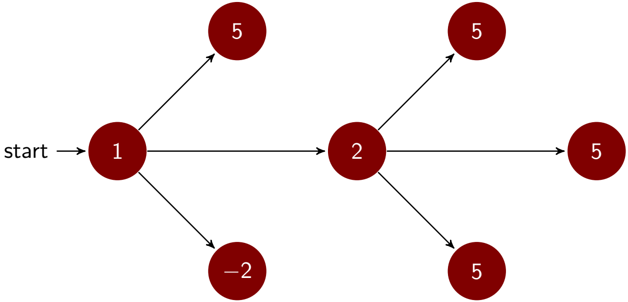
Summing up: 5.5. Going right was 6.9.

Time, risk and patience



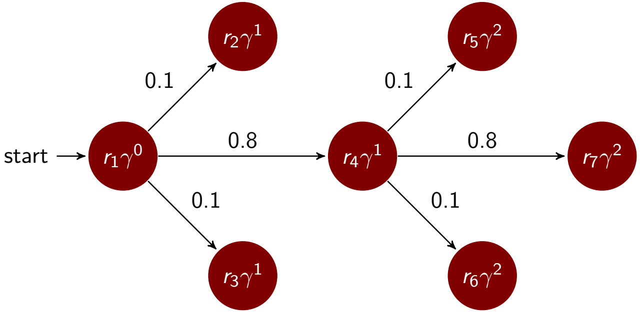
It's very easy to see that Down cannot be better than Up, which is already worse than Right.

Time, risk and patience



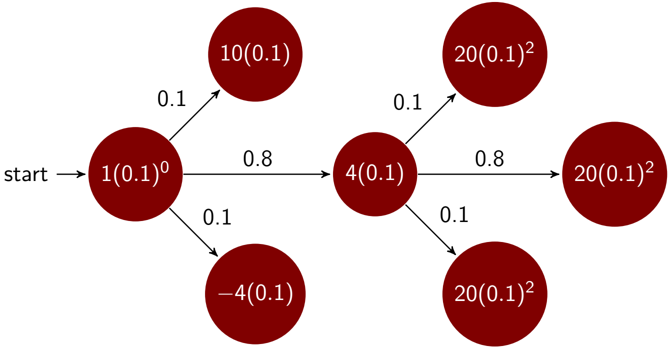
So going Right the first time is the best option.
And then any move works.

Time, risk and patience



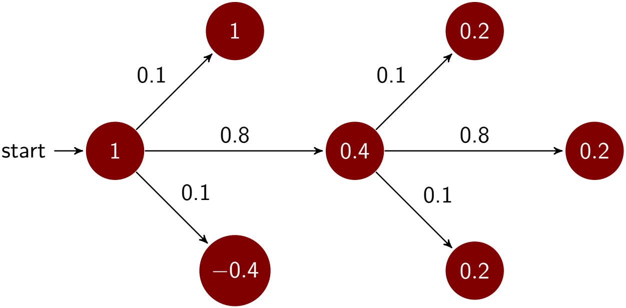
Remember that the best plan is determined by the rewards, the probabilities, and the discount factor.

Time, risk and patience



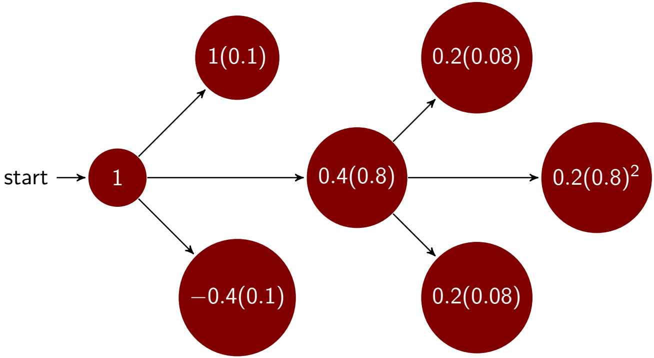
Suppose now $\gamma = 0.1$

Time, risk and patience



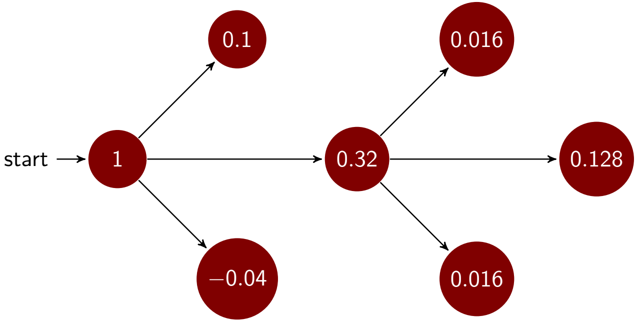
Notice how the perceived reward changes.

Time, risk and patience



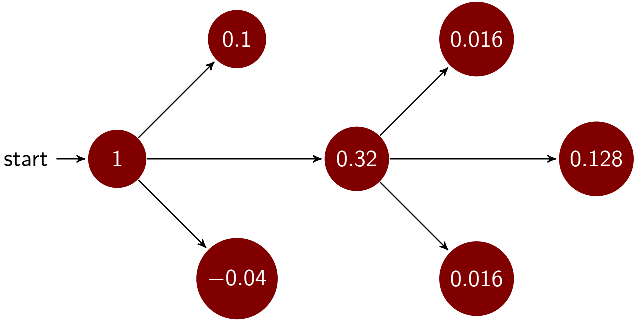
Including probabilities...

Time, risk and patience



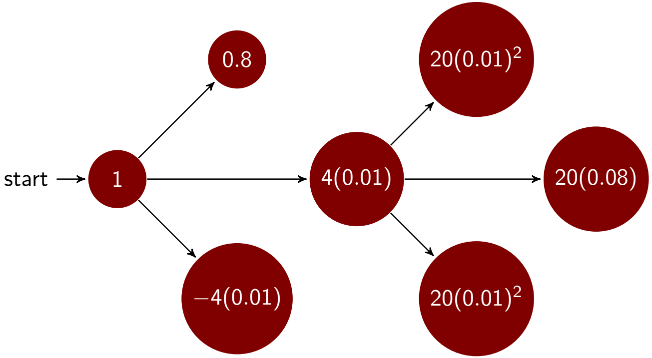
Notice the impact of discounting on negative rewards:
From very far away, all rewards look 0!

Time, risk and patience



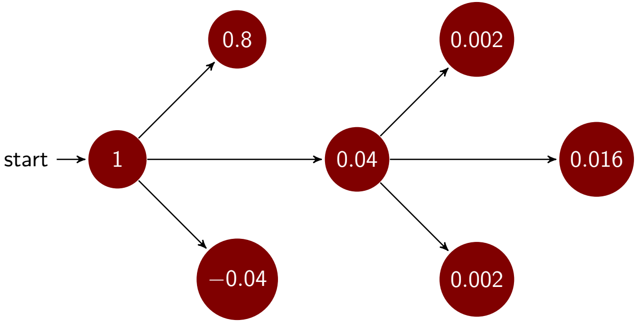
The expected utility at the starting state is: 1.54.

Time, risk and patience



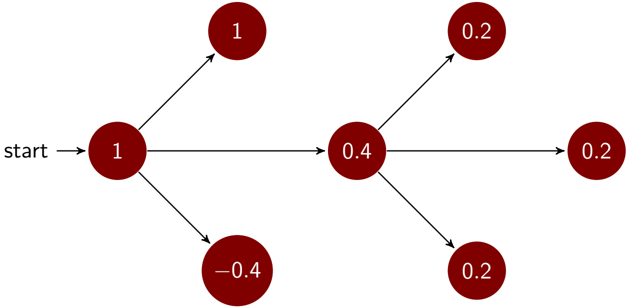
If we go Up instead...

Time, risk and patience



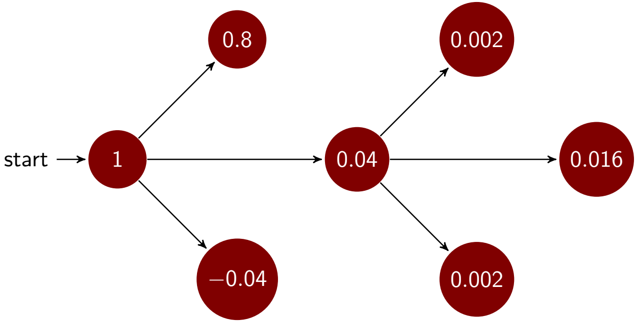
The expected utility at the starting state is: 1.82.
Going right was 1.54.

Time, risk and patience



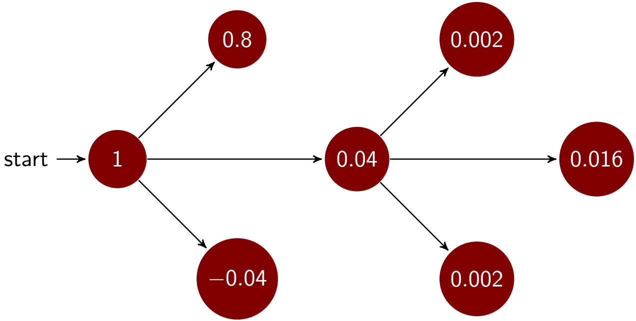
Again choosing Down is a poor choice.
In fact, unless $\gamma = 0$, it is always a poor choice.

Time, risk and patience



So, with $\gamma = 0.1$ any plan that goes *Up* first is the best choice.

Time, risk and patience



So, with $\gamma = 0.1$ any plan that goes *Up* first is the best choice.

Patience is key to decision-making

A problem

Here is a 3×101 world.

50	-1	-1	-1	$\cdots$	-1	-1	-1	-1
s				$\cdots$				
-50	1	1	1	$\cdots$	1	1	1	1

- start at s.
- two deterministic actions at s: either *Up* or *Down*
- beyond s you can only go *Right*.
- the numbers are the rewards you are going to get.

A problem

Here is a 3×101 world.

50	-1	-1	-1	$\cdots$	-1	-1	-1	-1
s				$\cdots$				
-50	1	1	1	$\cdots$	1	1	1	1

- start at s.
- two deterministic actions at s: either *Up* or *Down*
- beyond s you can only go *Right*.
- the numbers are the rewards you are going to get.

Compute the expected utility of each action as a function of γ

Solution

The utility of *Up* is

$$50\gamma - \sum_{t=2}^{101} \gamma^t = 50\gamma - \gamma^2 \frac{1-\gamma^{100}}{1-\gamma}$$

Solution

The utility of *Up* is

$$50\gamma - \sum_{t=2}^{101} \gamma^t = 50\gamma - \gamma^2 \frac{1-\gamma^{100}}{1-\gamma}$$

The utility of *Down* is

$$-50\gamma + \sum_{t=2}^{101} \gamma^t = -50\gamma + \gamma^2 \frac{1-\gamma^{100}}{1-\gamma}$$

Solution

The indifference point is

$$50\gamma - \gamma^2 \frac{1-\gamma^{100}}{1-\gamma} = -50\gamma + \gamma^2 \frac{1-\gamma^{100}}{1-\gamma}$$

Solution

The indifference point is

$$50\gamma - \gamma^2 \frac{1 - \gamma^{100}}{1 - \gamma} = -50\gamma + \gamma^2 \frac{1 - \gamma^{100}}{1 - \gamma}$$

Solving numerically, we have $\gamma \approx 0.9844$.

Solution

The indifference point is

$$50\gamma - \gamma^2 \frac{1-\gamma^{100}}{1-\gamma} = -50\gamma + \gamma^2 \frac{1-\gamma^{100}}{1-\gamma}$$

Solving numerically, we have $\gamma \approx 0.9844$.

- If γ is strictly larger than this then *Down* is better than *Up*;

Solution

The indifference point is

$$50\gamma - \gamma^2 \frac{1-\gamma^{100}}{1-\gamma} = -50\gamma + \gamma^2 \frac{1-\gamma^{100}}{1-\gamma}$$

Solving numerically, we have $\gamma \approx 0.9844$.

- If γ is strictly larger than this then *Down* is better than *Up*;
- If γ is strictly smaller than this then *Up* is better than *Down*;

Solution

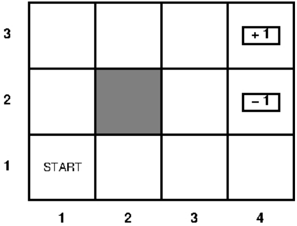
The indifference point is

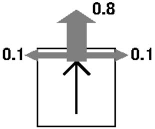
$$50\gamma - \gamma^2 \frac{1-\gamma^{100}}{1-\gamma} = -50\gamma + \gamma^2 \frac{1-\gamma^{100}}{1-\gamma}$$

Solving numerically, we have $\gamma \approx 0.9844$.

- If γ is strictly larger than this then *Down* is better than *Up*;
- If γ is strictly smaller than this then *Up* is better than *Down*;
- Else, it does not matter.

Makings plans





- The probability that $[Up, Up, Right, Right, Right]$ gets us to $+1$ is $0.32768 + 0.00008 = 0.32776$
- It does look like a really good plan: why so bad?

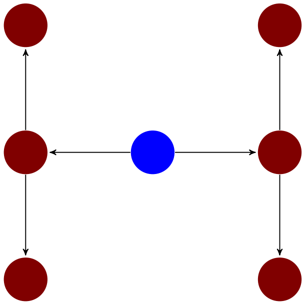
Plans vs Policies

- We have looked at a finite sequence of intended actions. But why would an agent stick to them even when the plan has failed?

Plans vs Policies

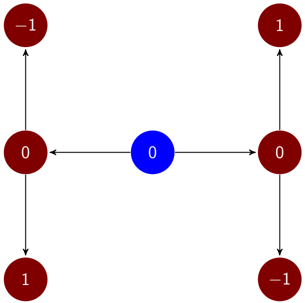
- We have looked at a finite sequence of intended actions. But why would an agent stick to them even when the plan has failed?
- The idea is that, if we know where we are, we need to think of what to do **there**.

An MDP



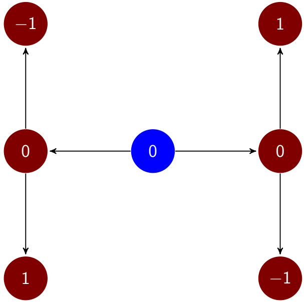
Start from the blue state.
First we go left or right, then up or down.

An MDP



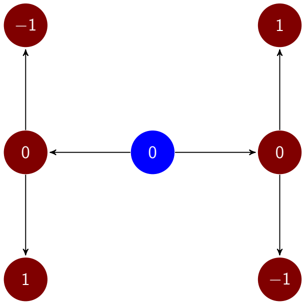
Let us assume the following rewards.
Forget about discounting.

An MDP



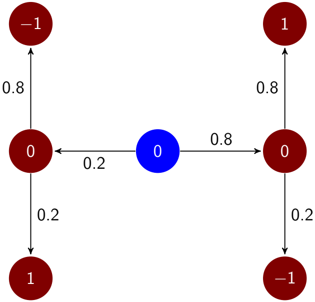
Assume 0.8 for the intended direction, 0.2 for the other.

An MDP



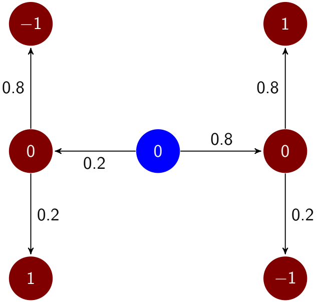
There are four possible plans:
[*Right*, *Up*], [*Right*, *Down*], [*Left*, *Up*], [*Left*, *Down*].
Let's check them out.

An MDP



$[Right, Up]$

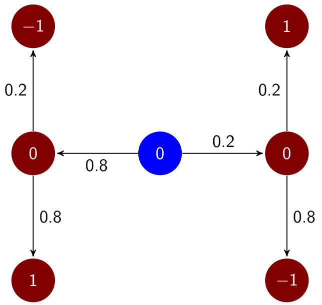
An MDP



$[Right, Up]$

The expected utility is $-2 \times 0.2 \times 0.8 + 0.8 \times 0.8 + 0.2 \times 0.2$

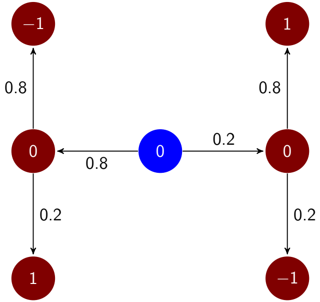
An MDP



[*Left*, *Down*]

The expected utility is $-2 \times 0.2 \times 0.8 + 0.8 \times 0.8 + 0.2 \times 0.2$

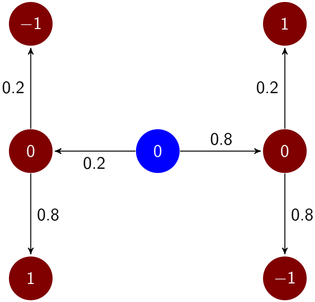
An MDP



[Left, Up]

The expected utility is $+2 \times 0.2 \times 0.8 - 0.8 \times 0.8 - 0.2 \times 0.2$

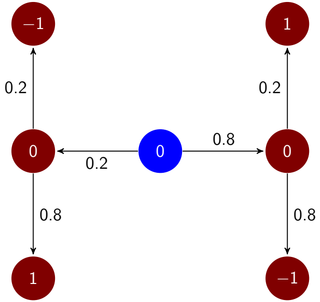
An MDP



[Right, Down]

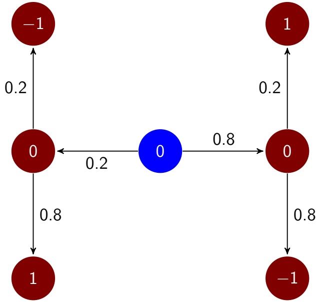
The expected utility is $+2 \times 0.2 \times 0.8 - 0.8 \times 0.8 - 0.2 \times 0.2$

An MDP



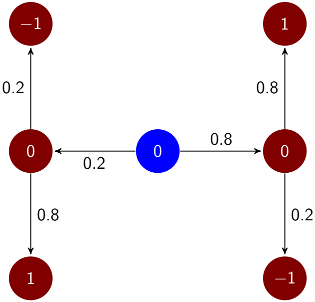
[*Right*, *Up*] and [*Left*, *Down*] are best, and we get
$-2 \times 0.2 \times 0.8 + 0.8 \times 0.8 + 0.2 \times 0.2$

An MDP



But can we really not be smarter than this?!

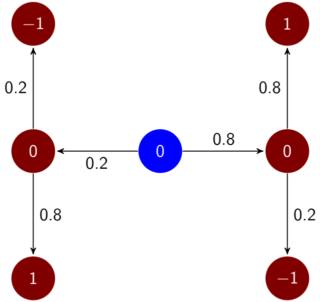
An MDP



What I want is:

- to go Down when I've gone Left!
- and to go Up when I've gone Right!

An MDP



The expected utility of this 'hybrid plan' is 0.6, higher than any plan we could possibly formulate.

Policies

Call S^+ the set of possible sequences of states. Call A the set of available actions.
Then a policy is a function:

$$S^+ \to 2^A \setminus \{\,\}$$

In words a policy is a protocol that at each possible decision moment prescribes a number of possible actions.

Policies

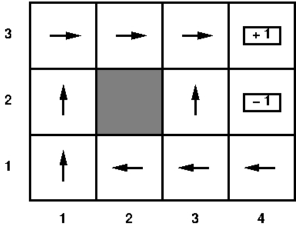
Call S^+ the set of possible sequences of states. Call A the set of available actions.
Then a policy is a function:

$$S^+ \to 2^A \setminus \{\}$$

In words a policy is a protocol that at each possible decision moment prescribes a number of possible actions.

- The intuition is that, according to that policy, at each stage we **should** perform one of the recommended actions.
- If multiple choices are recommended at some stage, then taking any of them means following the policy.

A policy



- This is a **state-based** policy. It recommends the same action at each state (so if two sequences end up with the same state, this policy is going to recommend the same action)
- This is a **deterministic** policy. At each state there is only one recommended action.

Expected utility of a policy

The expected utility (or value) of policy π, from state s is:

Expected utility of a policy

The expected utility (or value) of policy π, from state s is:

$$v^{\pi}(s) = E[\sum_{t=0}^{\infty} \gamma^{t} r(S_t)]$$

Expected utility of a policy

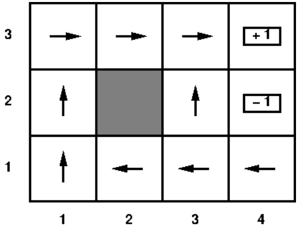
The expected utility (or value) of policy π, from state s is:

$$v^{\pi}(s) = E[\sum_{t=0}^{\infty} \gamma^t r(S_t)]$$

E is the expected utility of the sequences induced by:

- the policy π (the actions we are actually going to make)
- the initial state s (where we start)
- the transition model (where we can get to)

Loops



- In principle we can go on forever!
- We are going to assume we need to keep going unless we hit a terminal state (**infinite horizon assumption**)

Discounting

With discounting the utility of an infinite sequence is in fact **finite**.
If $\gamma < 1$ and rewards are bounded above by $\mathbf{r}$, we have:

$$u[s_1, s_2, \ldots] = \sum_{t=0}^{\infty} \gamma^t r(s_t) \leq \sum_{t=0}^{\infty} \gamma^t \mathbf{r} = \frac{\mathbf{r}}{1-\gamma}$$

Expected utility of a policy

An **optimal** policy (from a state) is the least deterministic [1] policy with the highest expected utility, starting from that state.

$$\pi_s^* = \underset{\pi}{\operatorname{argmax}}\, v^{\pi}(s)$$

We want to find the **optimal** policy.

[1] Least deterministic means that it always includes all the best actions, it there are more than one.

A remarkable fact

Theorem

With discounted rewards and infinite horizon

$$\pi^*_s = \pi^*_{s'}, \text{ for each } s' \in S$$

This means that the optimal policy does not depend on the sequences of states, but on the states only.
In other words, the optimal policy is a state-based policy.

A remarkable fact

Theorem

With discounted rewards and infinite horizon

$$\pi^*_s = \pi^*_{s'}, \text{ for each } s' \in S$$

This means that the optimal policy does not depend on the sequences of states, but on the states only.
In other words, the optimal policy is a state-based policy.

Idea: Take π^*_a and π^*_b. If they both reach a state c, because they are both optimal, there is no reason why they should disagree. So π^*_c is identical for both. But then they behave the same at all states!

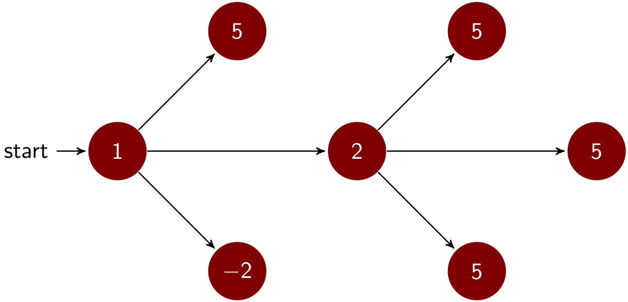
Value of states

The value of a state is the value of the optimal policy from that state.

In other words:

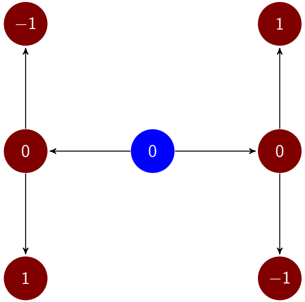
expected (discounted) sum of rewards
assuming optimal actions

Value of states



Assuming 0.8 to the intended direction, 0.1 otherwise...
The value of the starting state 6.9.

An MDP



Assuming 0.8 to the intended direction, 0.2 otherwise...
The value of the starting state is 0.6

VERY VERY IMPORTANT

Given the values of the states, choosing the best action is just maximisation of expected utility!

maximise the expected utility of the immediate successors

Value of states

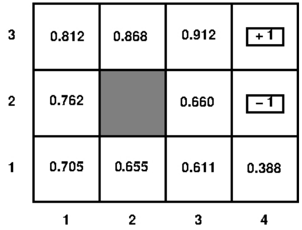


Figure : The values with $\gamma = 1$ and $r(s) = -0.04$

Value of states

	1	2	3	4
3	0.812	0.868	0.912	+1
2	0.762		0.660	−1
1	0.705	0.655	0.611	0.388

Figure : The optimal policy

$$\pi^*(s) = \underset{a \in A(s)}{\operatorname{argmax}} \sum_{s'} P(s' \mid s, a) v(s')$$

Maximise the expected utility of the subsequent state

Value of states

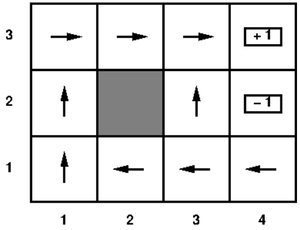


Figure : The optimal policy

$$\pi^*(s) = \underset{a \in A(s)}{\operatorname{argmax}} \sum_{s'} P(s' \mid s, a) v(s')$$

Maximise the expected utility of the subsequent state

Today's class

- Plans and policies

Coming next (but not next week, as there are no classes, nor tutorials, nor labs)

- How to find the optimal policy

Coming next (but not next week, as there are no classes, nor tutorials, nor labs)

- How to find the optimal policy