Planning and Optimization

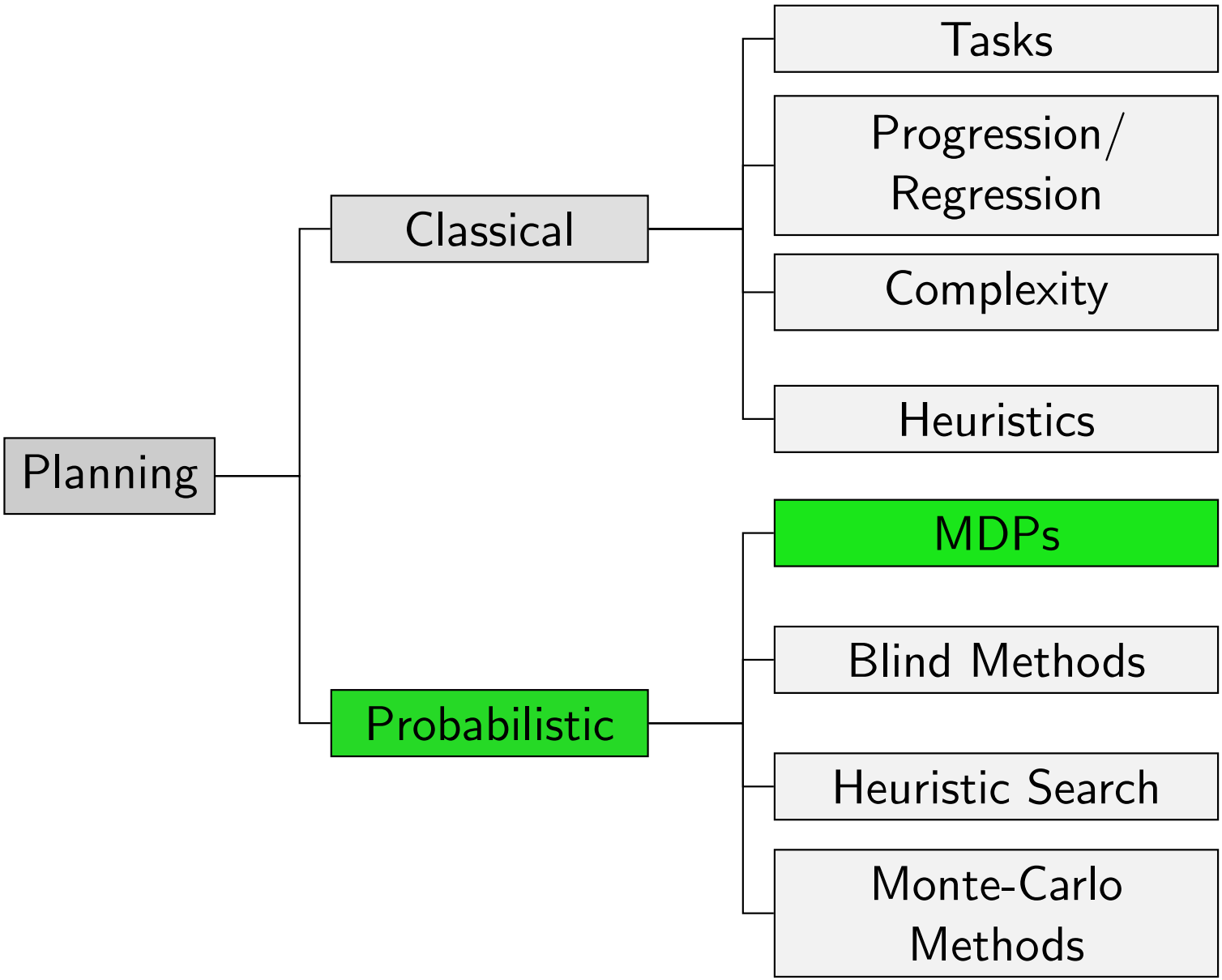
F1. Markov Decision Processes

Gabriele Röger and Thomas Keller

Universität Basel

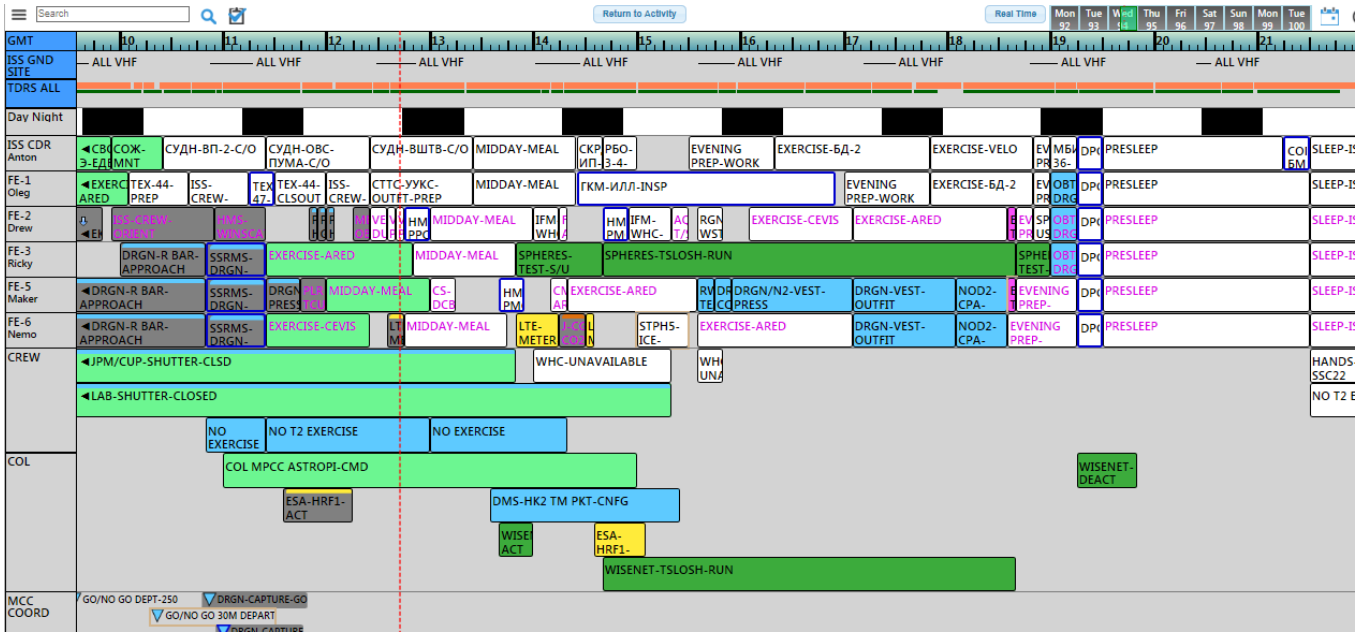
November 21, 2018

Content of this Course



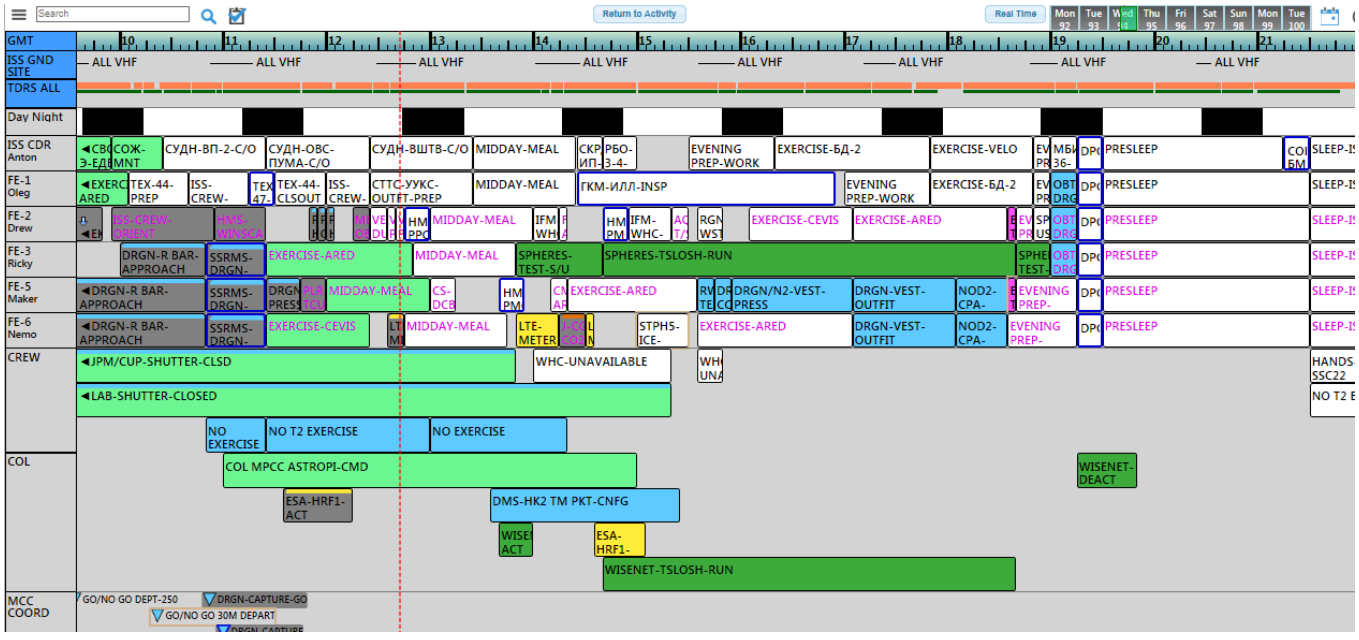
Motivation

Limitations of Classical Planning



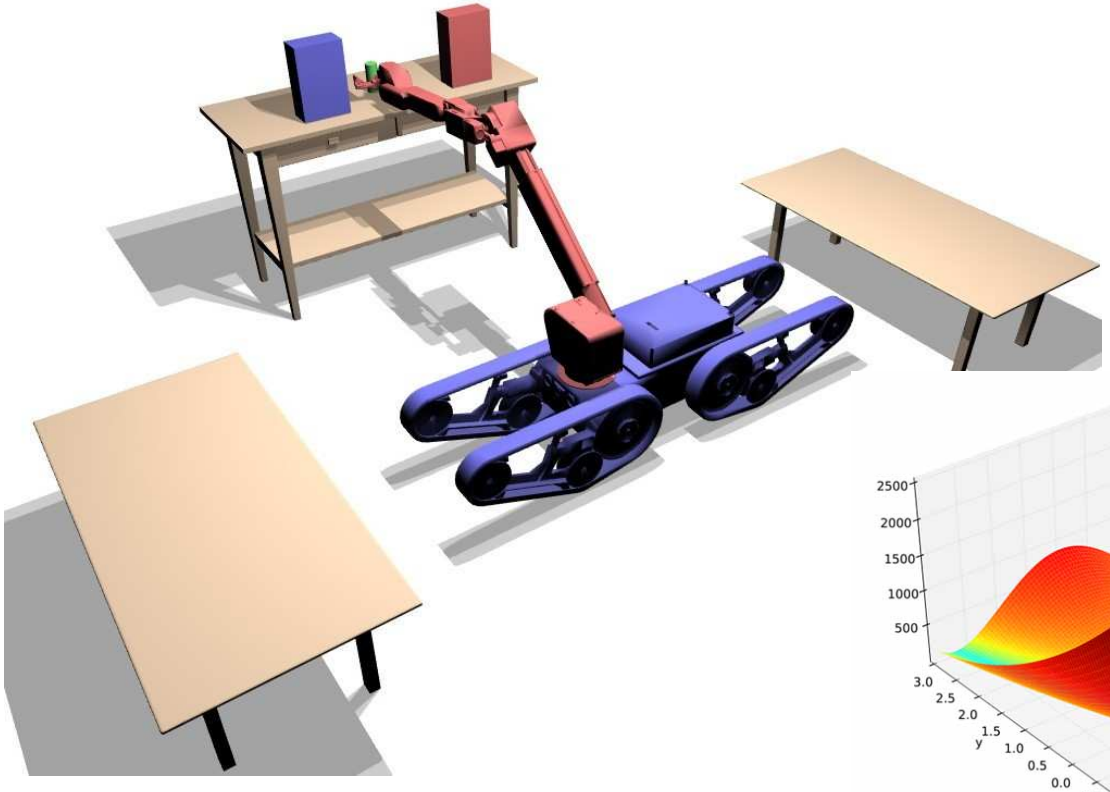
- timetable for astronauts on ISS

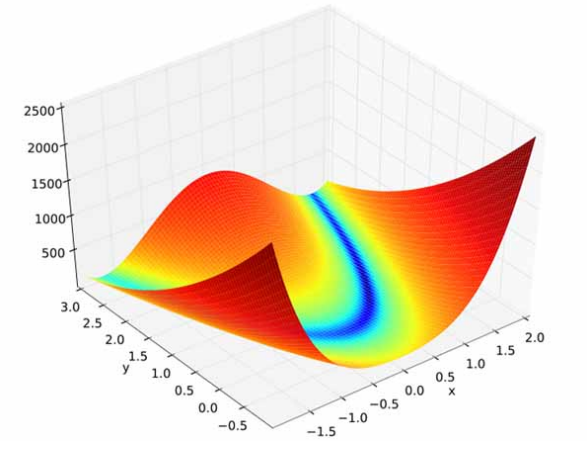
Generalization of Classical Planning: Temporal Planning



- timetable for astronauts on ISS
- concurrency required for some experiments
- optimize makespan

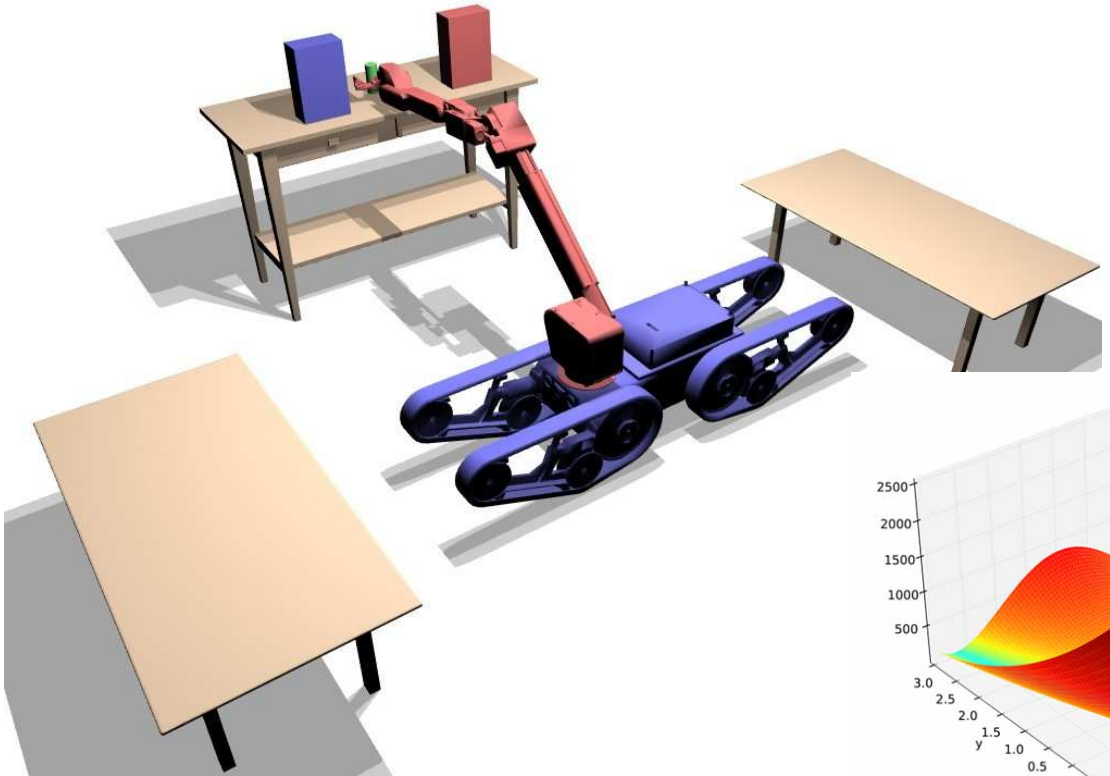
Limitations of Classical Planning

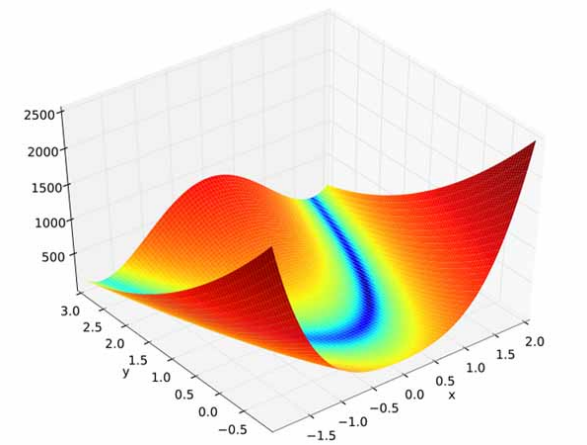




- kinematics of robotic arm

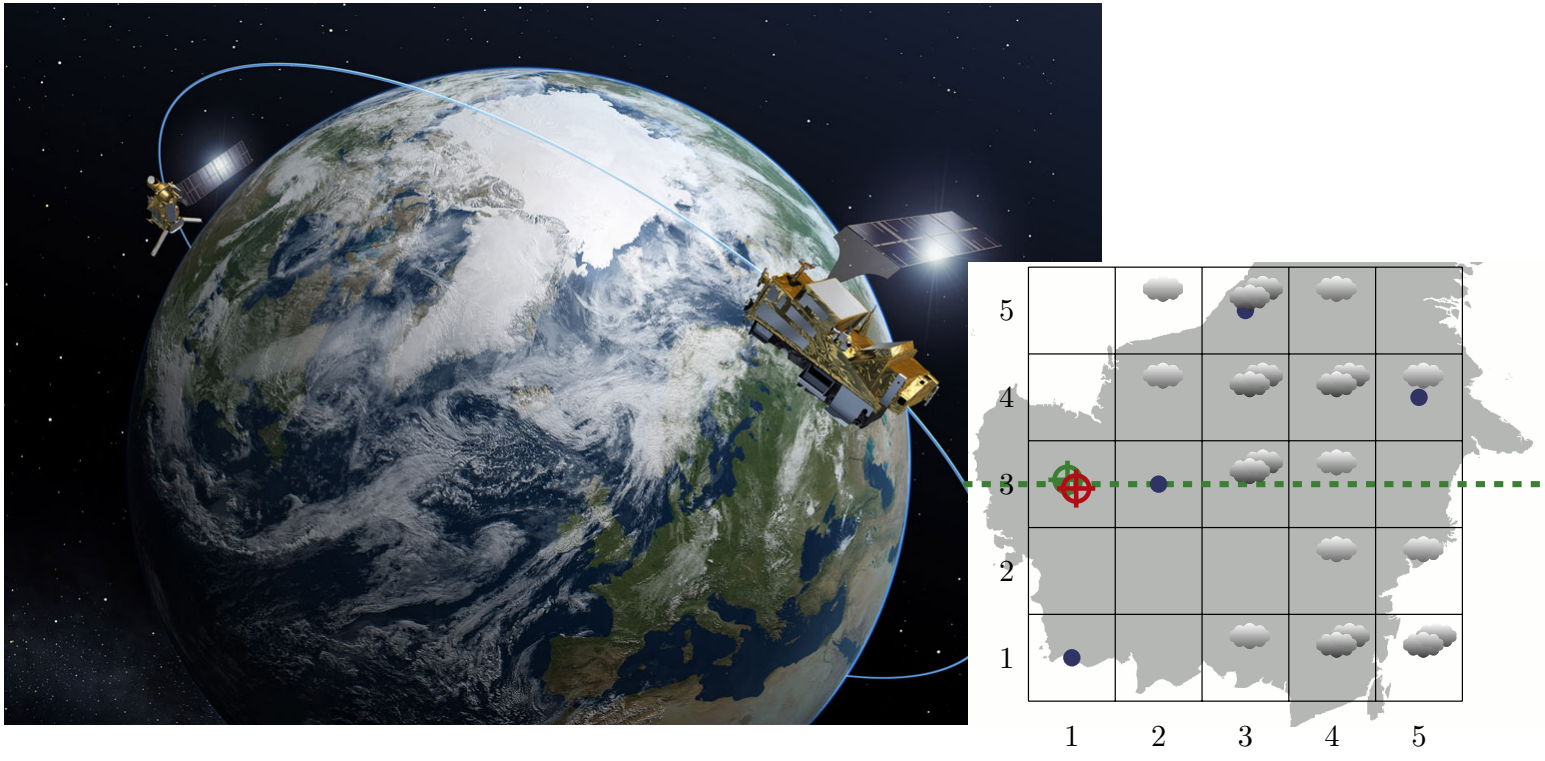
Generalization of Classical Planning: Numeric Planning





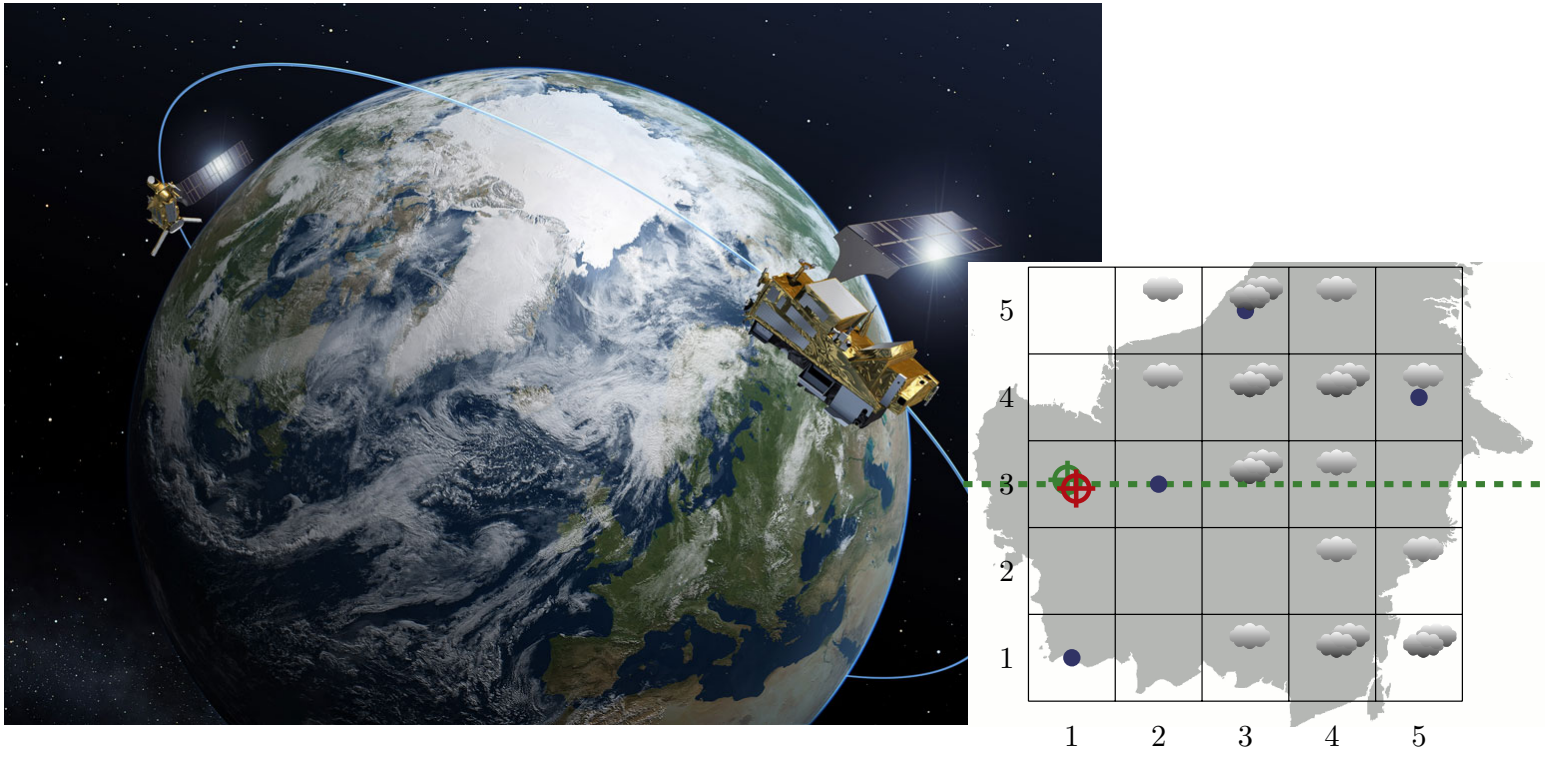
- kinematics of robotic arm
- state space is continuous
- preconditions and effects described by complex functions

Limitations of Classical Planning



- satellite takes images of patches on earth

Generalization of Classical Planning: MDPs



- satellite takes images of patches on earth
- weather forecast is uncertain
- find solution with lowest expected cost

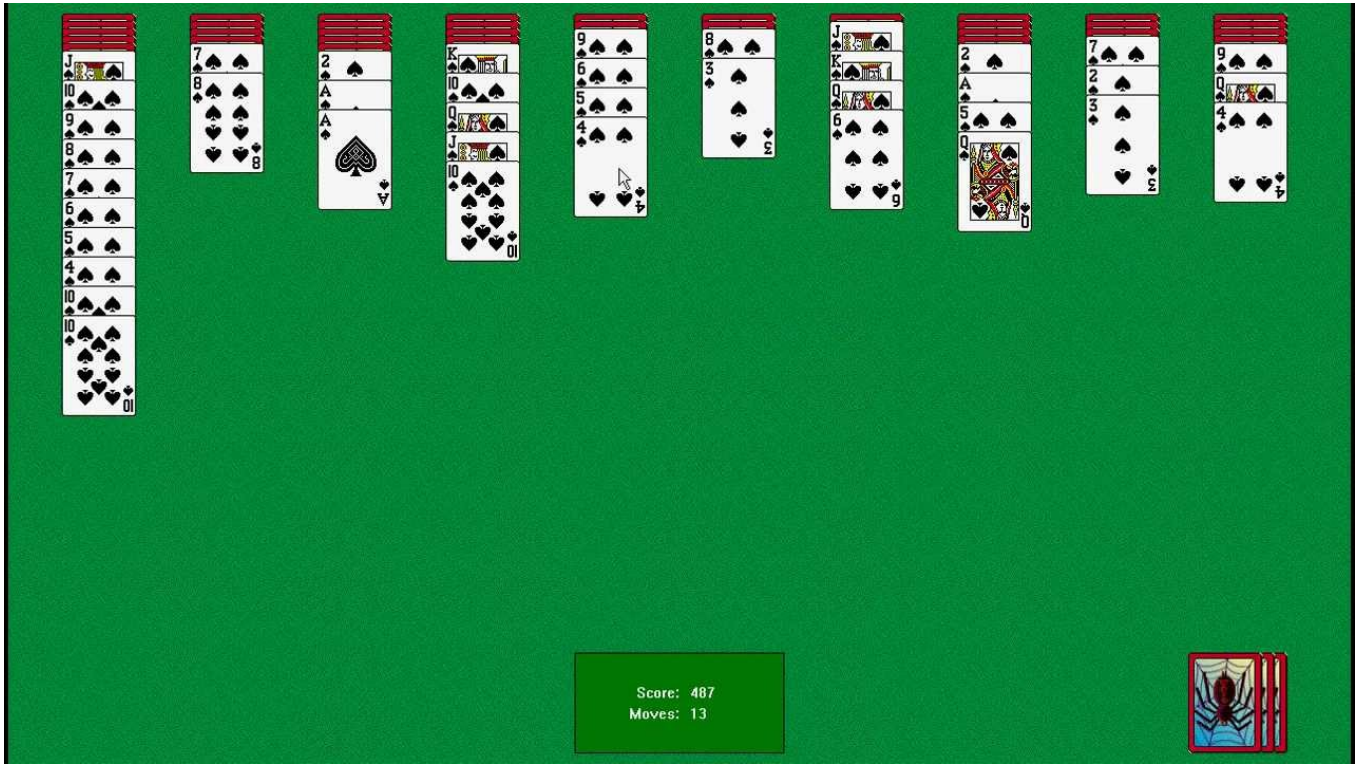
Limitations of Classical Planning

- Chess

Generalization of Classical Planning: Multiplayer Games

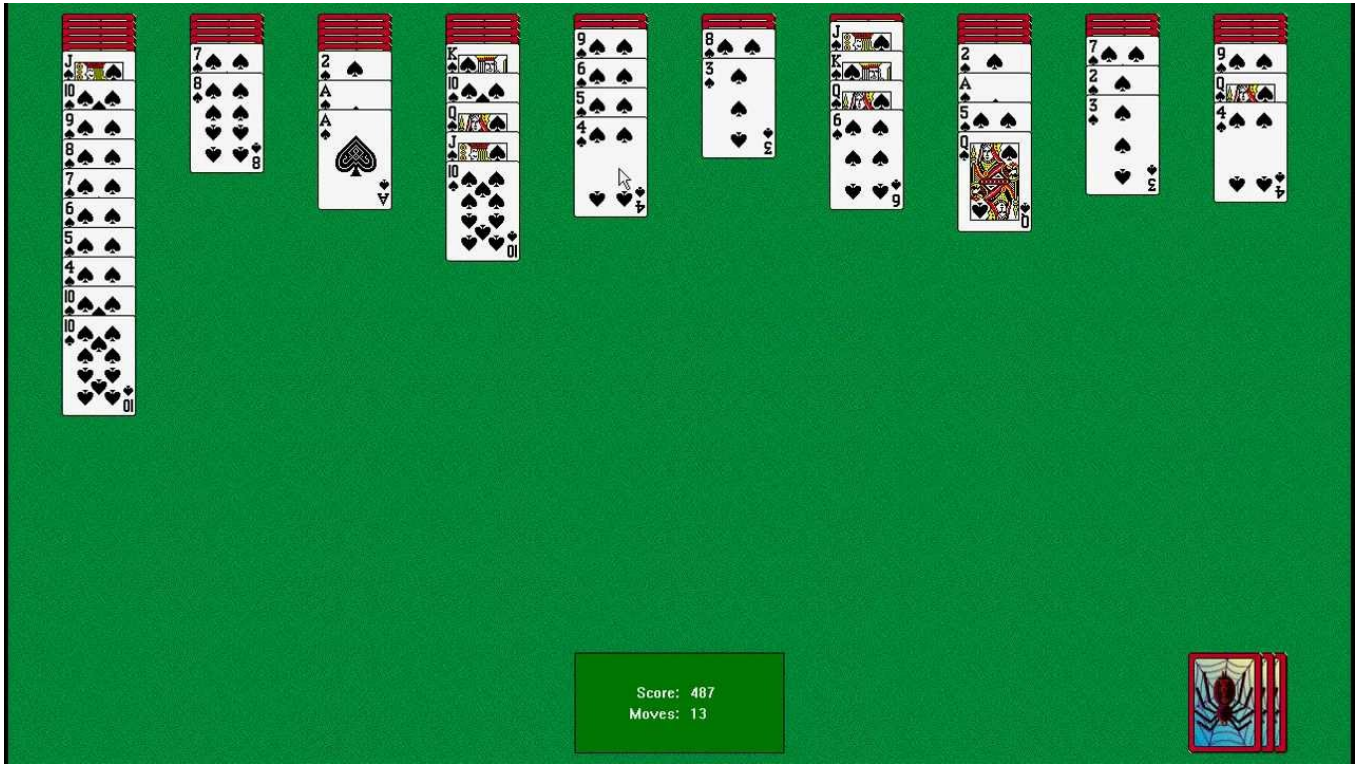
- Chess
- there is an opponent with
- contradictory objective

Limitations of Classical Planning



- Solitaire

Generalization of Classical Planning: POMDPs



- Solitaire
- some state information cannot be observed
- must reason over belief for good behaviour

Limitations of Classical Planning

- many applications are combinations of these
- all of these are active research areas
- we focus on one of them:
  probabilistic planning with Markov decision processes
- MDPs are closely related to games (Why?)

Markov Decision Processes

Markov Decision Processes

- Markov decision processes (MDPs) studied since the 1950s
- Work up to 1980s mostly on theory and basic algorithms for small to medium sized MDPs
- Today, focus on large (typically factored) MDPs
- Fundamental datastructure for reinforcement learning (not covered in this course)
- and for probabilistic planning
- different variants exist

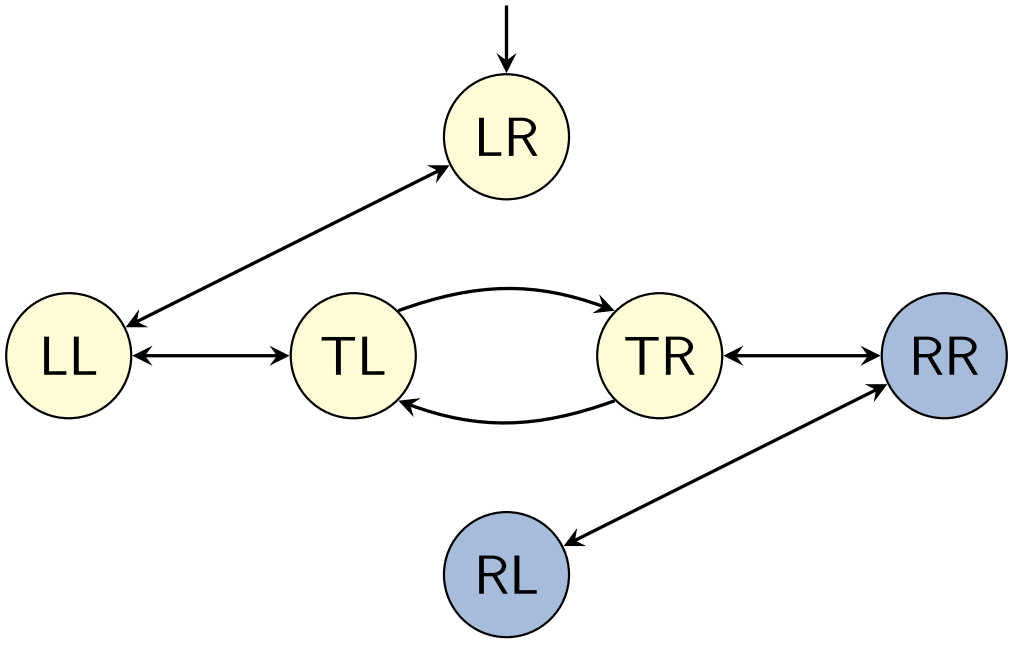
Reminder: Transition Systems

Definition (Transition System)

A transition system is a 6-tuple $\mathcal{T} = \langle S, L, c, T, s_0, S_\star \rangle$ where

- S is a finite set of states,
- L is a finite set of (transition) labels,
- $c : L \to \mathbb{R}_0^+$ is a label cost function,
- $T \subseteq S \times L \times S$ is the transition relation,
- $s_0 \in S$ is the initial state, and
- $S_\star \subseteq S$ is the set of goal states.

Reminder: Transition System Example



Logistics problem with one package, one trucks, two locations:

- location of package: $\{L, R, T\}$
- location of truck: $\{L, R\}$

Stochastic Shortest Path Problem

Definition (Stochastic Shortest Path Problem)

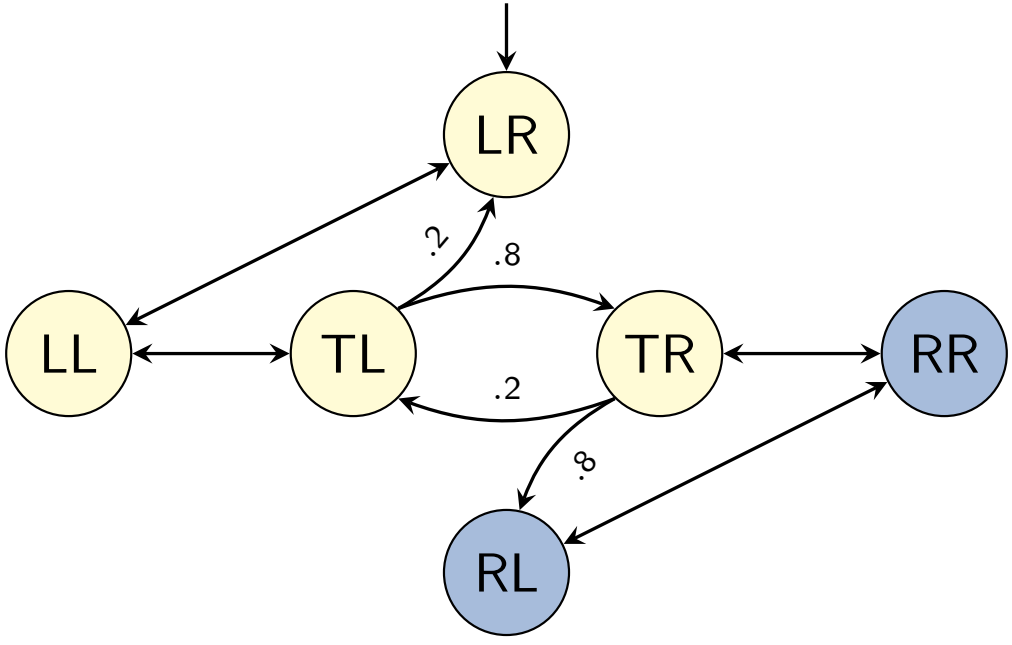
A stochastic shortest path problem (SSP) is a 6-tuple $\mathcal{T} = \langle S, L, c, T, s_0, S_\star \rangle$, where

- S is a finite set of states,
- L is a finite set of (transition) labels,
- $c : L \to \mathbb{R}_0^+$ is a label cost function,
- $T : S \times L \times S \mapsto [0, 1]$ is the transition function,
- $s_0 \in S$ is the initial state, and
- $S_\star \subseteq S$ is the set of goal states.

For all $s \in S$ and $\ell \in L$ with $T(s, \ell, s') > 0$ for some $s' \in S$, we require $\sum_{s' \in S} T(s, \ell, s') = 1$.

Note: An SSP is the probabilistic pendant of a transition system.

Reminder: Transition System Example



Logistics problem with one package, one trucks, two locations:

- location of package: $\{L, R, T\}$
- location of truck: $\{L, R\}$
- if truck moves with package, 20% chance of losing package

Terminology (1)

- If $p := T(s, \ell, s') > 0$, we write $s \xrightarrow{p:\ell} s'$ or $s \xrightarrow{p} s'$ if not interested in ℓ.
- If $T(s, \ell, s') = 1$, we also write $s \xrightarrow{\ell} s'$ or $s \rightarrow s'$ if not interested in ℓ.
- If $T(s, \ell, s') > 0$ for some s' we say that ℓ is applicable in s.
- The set of applicable labels in s is $L(s)$.

Terminology (2)

- the successor set of s and ℓ is $\text{succ}(s, \ell) = \{s' \in S \mid T(s, \ell, s') > 0\}$
- s' is a successor of s if $s' \in \text{succ}(s, \ell)$ for some ℓ
- s is predecessor of s' if $s' \in \text{succ}(s, \ell)$ for some ℓ
- with $s' \sim \text{succ}(s, \ell)$ we denote that successor $s' \in \text{succ}(s, \ell)$ of s and ℓ is sampled according to probability distribution T

Terminology (3)

- s' is reachable from s if there exists a sequence of transitions $s^0 \xrightarrow{p_1:\ell_1} s^1, \ldots, s^{n-1} \xrightarrow{p_n:\ell_n} s^n$ s.t. $s^0 = s$ and $s^n = s'$
  - Note: $n = 0$ possible; then $s = s'$
  - $s^0, \ldots, s^n$ is called (state) path from s to s'
  - $\ell_1, \ldots, \ell_n$ is called (label) path from s to s'
  - $s^0 \xrightarrow{\ell_1} s^1, \ldots, s^{n-1} \xrightarrow{\ell_n} s^n$ is called trace from s to s'
  - length of path/trace is n
  - cost of label path/trace is $\sum_{i=1}^{n} c(\ell_i)$
  - probability of path/trace is $\prod_{i=1}^{n} p_i$

Finite-horizon Markov Decision Process

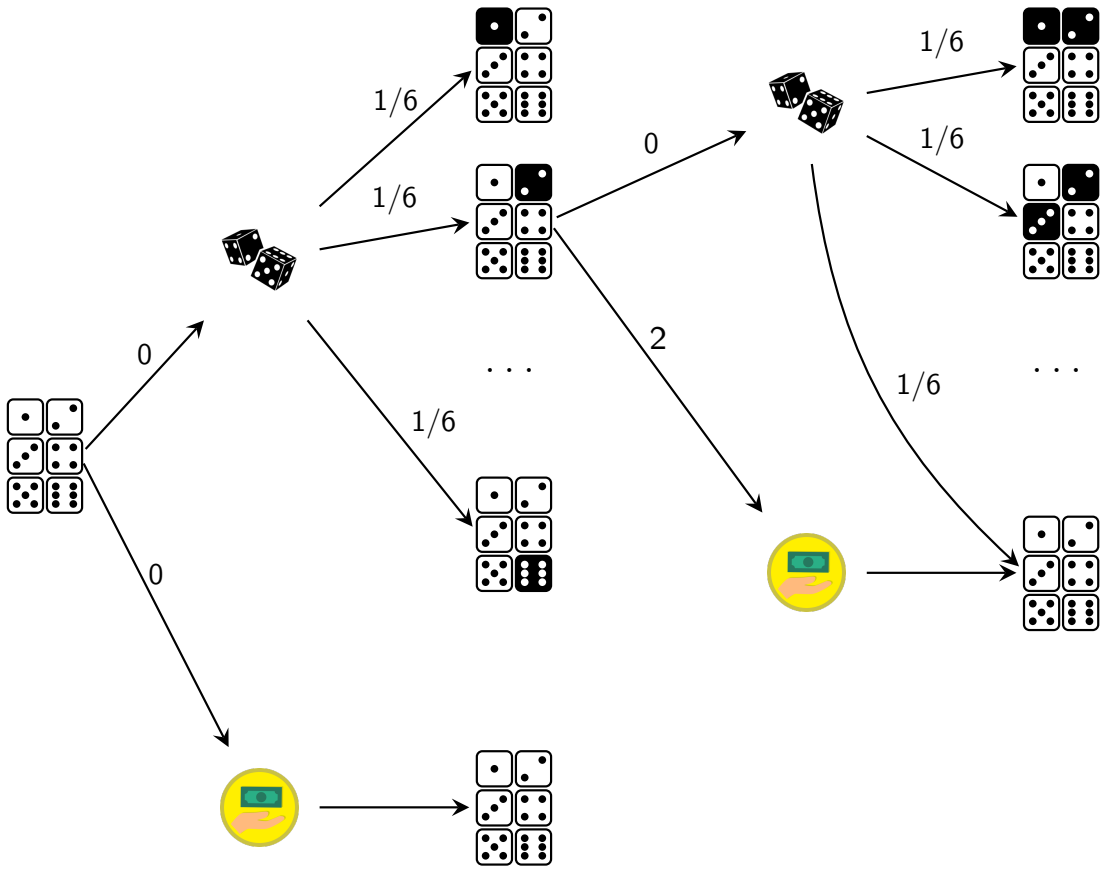
Definition (Finite-horizon Markov Decision Process)

A finite-horizon Markov decision process (FH-MDP) is a 6-tuple $\mathcal{T} = \langle S, L, R, T, s_0, H \rangle$, where

- S is a finite set of states,
- L is a finite set of (transition) labels,
- $R : S \times L \to \mathbb{R}$ is the reward function,
- $T : S \times L \times S \mapsto [0, 1]$ is the transition function,
- $s_0 \in S$ is the initial state, and
- $H \in \mathbb{N}$ is the finite horizon.

For all $s \in S$ and $\ell \in L$ with $T(s, \ell, s') > 0$ for some $s' \in S$, we require $\sum_{s' \in S} T(s, \ell, s') = 1$.

Example: Push Your Luck



Discounted Reward Markov Decision Process

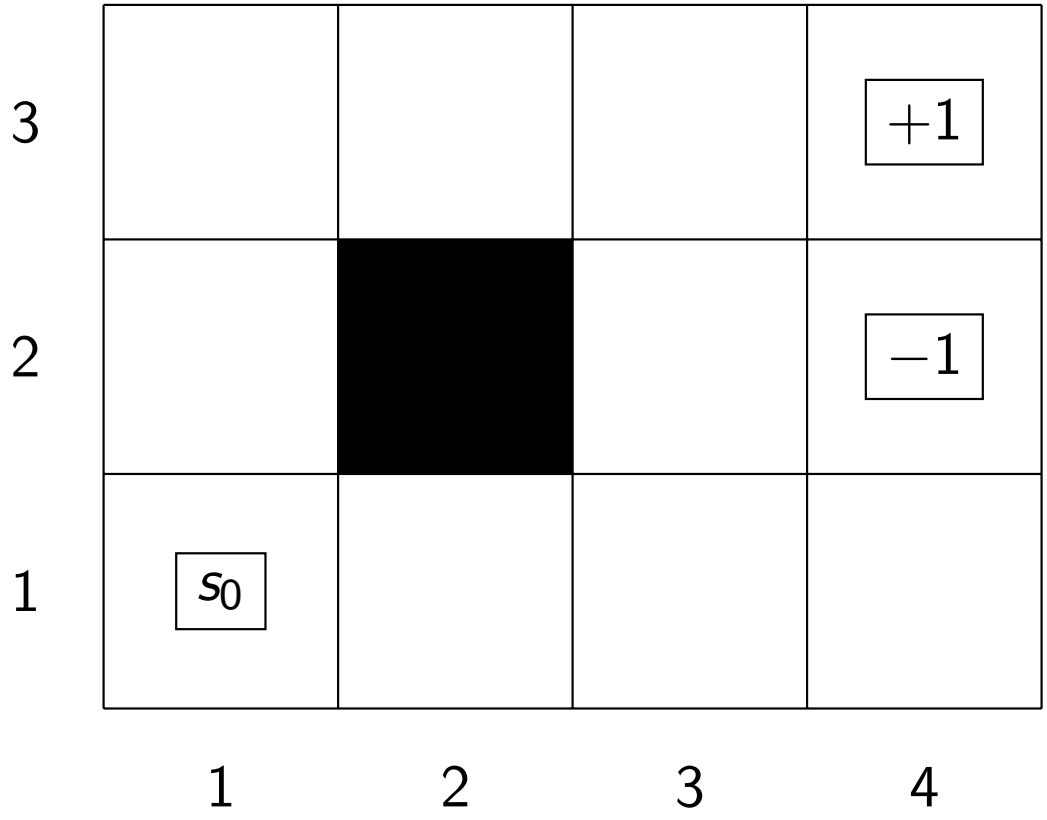
Definition (Discounted Reward Markov Decision Process)

A discounted reward Markov decision process (DR-MDP) is a 6-tuple $\mathcal{T} = \langle S, L, R, T, s_0, \gamma \rangle$, where

- S is a finite set of states,
- L is a finite set of (transition) labels,
- $R : S \times L \to \mathbb{R}$ is the reward function,
- $T : S \times L \times S \mapsto [0, 1]$ is the transition function,
- $s_0 \in S$ is the initial state, and
- $\gamma \in (0, 1)$ is the discount factor.

For all $s \in S$ and $\ell \in L$ with $T(s, \ell, s') > 0$ for some $s' \in S$, we require $\sum_{s' \in S} T(s, \ell, s') = 1$.

Example: Grid World



- each move goes in orthogonal direction with some probability
- (4,3) gives reward of $+1$ and sets position back to (1,1)
- (4,2) gives reward of -1

Summary

Summary

- Many planning scenarios beyond classical planning
- We focus on probabilistic planning
- SSPs are classical planning + probabilistic transition function
- FH-MDPs and DR-MDPs allow state-dependent rewards
- FH-MDPs consider finite number of steps
- DR-MDPs discount rewards over infinite horizon