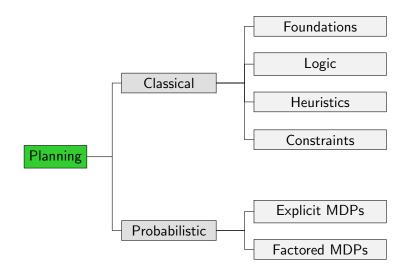
Planning and Optimization F1. Markov Decision Processes

Malte Helmert and Gabriele Röger

Universität Basel

Stochastic Shortest Path Problem

#### Content of this Course



Stochastic Shortest Path Problem

Summary 00

# **Motivation**

Stochastic Shortest Path Problem 000000000

Summary 00

#### Limitations of Classical Planning

E Search		۵ 🕅							Pal	tern to Act	Ivly.						Real To		1	Heat The	fri Sut Sun Mi UT Mi	Man Tar 🛗
GMT ISS GND SITE	- ALL VHF		ALL VHP	h2, 1,	44	ALL VHF	dute	. 14. 1	-AU W	HF LS	ш	<u></u> A	1. VHF	hutu		18.1. M7	du.	A	LL VH		- ALL VH	8.1.1.1.1.
TORS ALL Day Night																						
ISS CDR Anton	«Свесож- судн-	8 <b>П-2-</b> С/О	СУДН-ОВ	с. 0	судн	BUTB-C/O	MIDDAY	MEAL	OKP PE	50- 4-		EVENING PREP-WOR	x D	KERCISE-E	ф-2	EXERCISE	VELO	EV ME PR 36-	POP	PRESLEEP		COI SLEEP-
FE-1 Oleg	ARED PREP CR	5. T IEW. 4	TEX-44-	ISS- CREW-	CTTC:	YYKC- I-PREP	MIDDAY	-MEAL	CKM-P	клл-ans	iP				EVENING PREP-WORK	EXERCISE	6Д-2	EV 08 PR DR	0P	PRESLEEP		SLEEP-
FE-2 Drew	AD DRENT	HMS- WD/SC			VE V	HM MIDDA	AY-MEAL	3FM WH	H	M PM-	. #	RGN	EXERC	ISE-CEVIS	EXERCISE-AR	£D		SP DE	07	PRESLEEP		SLEEP-
FE-3 Ricky	DRGN-R BAR-	SSRMS-	EXERCISE	-ARED		MIDDAY-	MEAL	SPHERE TEST-S/	s. 19	PHERES	TSLOS	H-RUN					ł	PHE OS	0P	PRESLEEP		SLEEP-
FE-5 Maker	ADRGN-R BAR-	SSRMS- DRGN-	DRGN - I	MIDDA	o-mb	L (S-	HM		EXERCIS	E-ARED		RV DF DF TE CC PR	GN/N2 ESS	-VEST-	DRGN-VEST-	NOD CPA-		VENING	0P	PRESLEEP		SLEEP-
FE-6 Nemo	DRGN-R BAR- APPROACH	SSRMS-	EXERCISE	-CEVIS	g	MIDDAY-N		TE-		STP	HS-	EXERCIS			DRGN-VEST-	NOD	2- EV PR	ENING SP:	OP	PRESLEEP		SLEEP-
CREW	JPM/CUP-SHUTTER-	CLSD			T			WHO	UNAVAI	LABLE		WH										HAND
	<lab-shutter-closi< td=""><td>ED</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>NO T2</td></lab-shutter-closi<>	ED																				NO T2
		NO	NO T2 EX	ERCISE		NO EX	ERCISE				_											
COL	1	COLN	APCC ASTRO	OND-CWD															WES DEA	ENET-		
	ESA.H3F1-						DMS-HK2 TM PKT-CNFG												_			
				_			W2S ACT		ESA HRU	ь 11.	_											
							WISENET-TSLOSH-RUN															
MCC COORD	GO/NO GO DEPT-250																					

#### timetable for astronauts on ISS

Stochastic Shortest Path Problem

## Generalization of Classical Planning: Temporal Planning

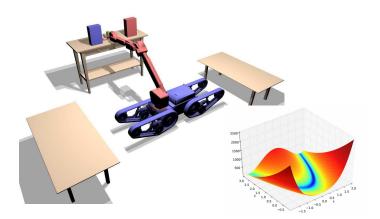
E Search		۵ 🕅						10	dam to Asthety					10	6 7294	Mar	12	Hed The Fri St	d San Mon	
GMT ISS GND SITE	- ALL VHT		ALL VHF	<u>-</u>	ALL VHF	dute	44.1	- AU V	ни ни		ALL VIE		UL VH	, <mark>48, 1, , 1</mark> , 7	-	, 19, AL	L VH		ALL 197	ulute
TDRS ALL Day Night				-				_		-							-			
ISS CDR Anton	«свесож. судн.	80-2-6/0	СУДН-ОВС- ПУМА-С/О	суді	-BWTB-C/O	MIDDAY	MEAL	CKP P	FO-	EVENIN PREP-W	5 EXERCI	5Е-БД-2	E	DERCISE-VEL	0	EV MEA	OP	PRESLEEP		COL SLEEP.
FE-1 Oleg	EXERCITEX-44- IS	5. IEW.	TEX TEX-44- ISS-	CTTC	YYKC- T-PREP	MIDDAY	MEAL		илл-эмэр	(FALL IN		EVENING PREP-WOR	ж <sup>Е</sup>	XERCISE-64			OP	PRESLEEP		SLEEP-
FE-2 Drew	AD DRENT	HMS-		VEN DUP	HM MIDDA	Y-MEAL	SFM F WHO		M BFM- M WHC-	RGN	EXERCISE-C	EVIS EXERCISE	-AREC	>	£5 29	SP DAT	07	PRESLEEP		SLEEP-
FE-3 Ricky	DRGN-R BAR- APPROACH	SSRMS- DRGN	EXERCISE-ARED		MIDDAY-	MEAL	PHERES EST-S/L	5	PHERES-TSU	DSH-RUN					SPI TES	DRC	0PK	PRESLEEP		SLEEP
FE-5 Maker	ADRGN-R BAR- APPROACH	SSRMS- DRGN-	PRES TEL	er-ME	u CS- DCB	HM PM	C1 A3	DURCE	HE-ARED	RI DA TE CO	ORGN/N2-VES PRESS	DRGN-V OUTPIT	151-	NOD2- CPA-	EVI 281	NING 9-	0PK	PRESLEEP		SLEEP-
FE-6 Nemo	<pre> ADRGN-R BAR- APPROACH </pre>	SSRMS- DRGN-	EXERCISE-CEVIS	L.	MIDDAY-M	EAL	TE- IETER		STPHS- KE-	EXERC	SE-ARED	DRGN-V OUTFIT	EST-	NOD2- CPA-	EVEN PREP	ang .	OP	PRESLEEP		SLEEP-I
CREW	JPM/CUP-SHUTTER-						WHC	-UNAVA	LABLE	WH										HANDS
	<lab-shutter-clos< td=""><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>NO T2</td></lab-shutter-clos<>																			NO T2
COL	_	NO EXERCIS			NO EX	ERCISE			_								_			
coc		COL	MPCC ASTROPI-CME	<u> </u>													DEA	ENET- CT		
			ESA-HRF1- ACT			-		PKT-CN												
						W2S ACT			F1-											
MCC			APTLEE-GO					. (	risene1-15L	usm-RUN	_	_								
COORD	¥00/N0	SO SOM DE	PART																	

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

Markov Decision Proces: 00000000000 Stochastic Shortest Path Problem

Summary 00

## Limitations of Classical Planning

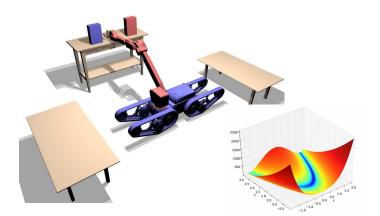


kinematics of robotic arm

Markov Decision Process

Stochastic Shortest Path Problem

## Generalization of Classical Planning: Numeric Planning

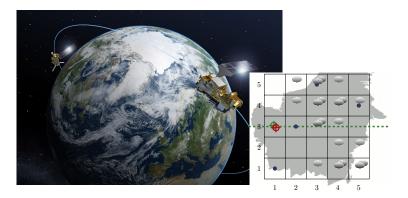


- kinematics of robotic arm
- state space is continuous
- preconditions and effects described by complex functions

Markov Decision Process 000000000000 Stochastic Shortest Path Problem

Summary 00

## Limitations of Classical Planning

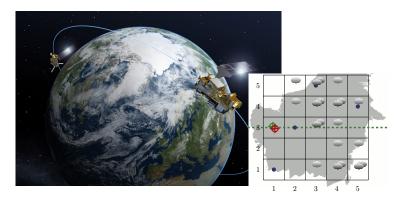


satellite takes images of patches on earth

Markov Decision Process 00000000000 Stochastic Shortest Path Problem

Summary 00

#### Generalization of Classical Planning: MDPs



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

Markov Decision Process 000000000000 Stochastic Shortest Path Problem

Summary 00

#### Limitations of Classical Planning





Markov Decision Process 00000000000 Stochastic Shortest Path Problem

### Generalization of Classical Planning: Multiplayer Games



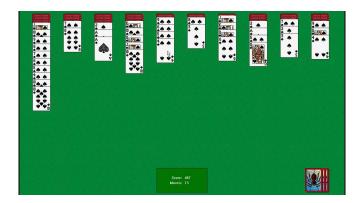
#### Chess

there is an opponent with a contradictory objective

Markov Decision Proces: 000000000000 Stochastic Shortest Path Problem

Summary 00

## Limitations of Classical Planning





Markov Decision Process 00000000000 Stochastic Shortest Path Problem

Summary 00

#### Generalization of Classical Planning: POMDPs



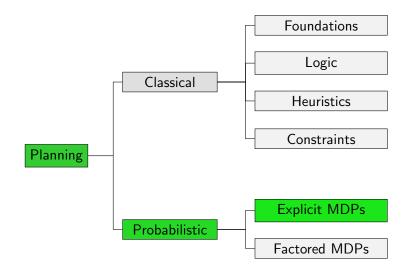
- Solitaire
- some state information cannot be observed
- must reason over belief for good behaviour

# 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?)

Stochastic Shortest Path Problem

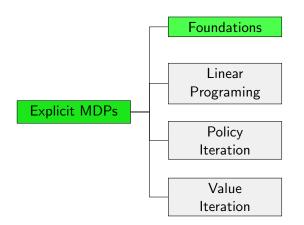
#### Content of this Course



Stochastic Shortest Path Problem

Summary 00

### Content of this Course: Explicit MDPs



Stochastic Shortest Path Problem

# Markov Decision Process

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

Stochastic Shortest Path Problem 000000000

# 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.

 $\rightarrow$  goal states and deterministic transition function

#### Definition (Markov Decision Process)

- A (discounted reward) Markov decision process (MDP) is a 6-tuple  $\mathcal{T} = \langle S, A, R, T, s_0, \gamma \rangle$ , where
  - S is a finite set of states,
  - A is a finite set of actions,
  - $R: S \times A \rightarrow \mathbb{R}$  is the reward function,
  - $T: S \times A \times S \mapsto [0, 1]$  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  $a \in A$  with T(s, a, s') > 0 for some  $s' \in S$ , we require  $\sum_{s' \in S} T(s, a, s') = 1$ .

## Reward instead of Goal States

- the agent does not try to reach a goal state but gets a (positive or negative) reward for each action application.
- infinite horizon: agent acts forever
- finite horizon: agent acts for a specified number of steps
- we only consider the variant with an infinite horizon
- immediate reward is worth more than later reward
  - as in economic investments
  - ensures that our algorithms will converge
- $\blacksquare$  the value of a reward decays exponentially with  $\gamma$
- now full value r, in next step  $\gamma r$ , in two steps only  $\gamma^2 r$ , ...
- aim: maximize expected overall reward

Stochastic Shortest Path Problem

Summary 00

## Markov Property

Why is this called a Markov decision process?

Russian mathematician Andrey Markov (1856–1922)

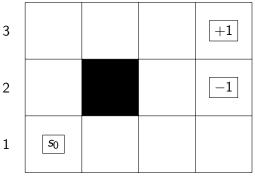


Markov property: the probability distribution for the next state only depends on the current state (and the action) but not on previously visited states or earlier actions.

Markov Decision Process

Stochastic Shortest Path Problem

# Example: Grid World



1 2 3 4

- moving north goes east with probability 0.4
- only applicable action in (4,2) and (4,3) is *collect*, which
  - sets position back to (1,1)
  - gives reward of +1 in (4,3)
  - gives reward of -1 in (4,2)

# Solutions in MDPs

#### classical planning

- a solution is a sequence of operators
- next state always clear
- at the end we are in a goal state
- MDP
  - next state uncertain
  - we cannot know in advance what actions will be applicable in the encountered state
  - infinite horizon: act forever
  - $\blacksquare$   $\rightarrow$  sequence of operators does not work
  - $\blacksquare$   $\rightarrow$  policy: specify for each state the action to take
  - $\blacksquare$   $\rightarrow$  at least for all states which we can potentially reach

Stochastic Shortest Path Problem 000000000

# Terminology (1)

- If p := T(s, a, s') > 0, we write  $s \xrightarrow{p:a} s'$ (or  $s \xrightarrow{p} s'$  if a is not relevant).
- If T(s, a, s') = 1, we also write  $s \xrightarrow{a} s'$  or  $s \rightarrow s'$ .
- If T(s, a, s') > 0 for some s' we say that a is applicable in s.
- The set of applicable actions in s is A(s). We assume that  $A(s) \neq \emptyset$  for all  $s \in S$ .

# Terminology (2)

- the successor set of s and a is  $\operatorname{succ}(s, a) = \{s' \in S \mid T(s, a, s') > 0\}.$
- s' is a successor of s if  $s' \in \text{succ}(s, a)$  for some a.
- to indicate that s' is a successor of s and a that is sampled according to probability distribution T, we write s' ~ succ(s, a)

Stochastic Shortest Path Problem

# Policy for MDPs

#### Definition (Policy for MDPs)

Let  $\mathcal{T} = \langle S, A, R, T, s_0, \gamma \rangle$  be a (discounted-reward) MDP. Let  $\pi$  be a mapping  $\pi : S \to A \cup \{\bot\}$  such that  $\pi(s) \in A(s) \cup \{\bot\}$  for all  $s \in S$ .

The set of reachable states  $S_{\pi}(s)$  from s under  $\pi$  is defined recursively as the smallest set satisfying the rules

• 
$$s \in S_{\pi}(s)$$
 and

•  $\operatorname{succ}(s', \pi(s')) \subseteq S_{\pi}(s)$  for all  $s' \in S_{\pi}(s)$  where  $\pi(s') \neq \bot$ .

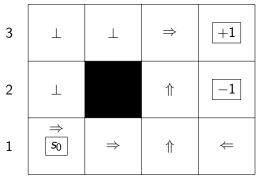
If  $\pi(s') \neq \bot$  for all  $s' \in S_{\pi}(s_0)$ , then  $\pi$  is a policy for  $\mathcal{T}$ .

Markov Decision Process

Stochastic Shortest Path Problem

Summary 00

# Example: Grid World



1 2 3 4

- moving north goes east with probability 0.4
- only applicable action in (4,2) and (4,3) is *collect*, which
  - sets position back to (1,1)
  - gives reward of +1 in (4,3)
  - sives reward of -1 in (4,2)

Stochastic Shortest Path Problem

# Stochastic Shortest Path Problem

#### I Want My Goal States Back!

- We also consider a variant of MDPs that are not discounted-reward MDPs.
- Stochastic Shortest Path Problems (SSPs) are closer to classical planning.
  - goal states
  - but still stochastic transition function
- We will use the same concepts for SSPs as for discounted-reward MDPs (e.g. policies)

# Stochastic Shortest Path Problem

Definition (Stochastic Shortest Path Problem)

A stochastic shortest path problem (SSP) is a 6-tuple

- $\mathcal{T} = \langle \mathcal{S}, \mathcal{A}, \mathcal{c}, \mathcal{T}, \mathcal{s}_0, \mathcal{S}_\star 
  angle$ , where
  - *S* is a finite set of states,
  - A is a finite set of actions,
  - $c: A \to \mathbb{R}^+_0$  is an action cost function,
  - $T: S \times A \times S \mapsto [0, 1]$  is the transition function,
  - $s_0 \in S$  is the initial state, and
  - $S_{\star} \subseteq S$  is the set of goal states.

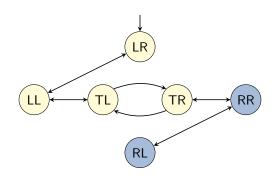
For all  $s \in S$  and  $a \in A$  with T(s, a, s') > 0 for some  $s' \in S$ , we require  $\sum_{s' \in S} T(s, a, s') = 1$ .

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

Stochastic Shortest Path Problem

Summary 00

#### Transition System Example



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

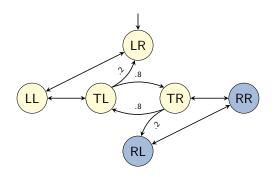
- location of package: domain  $\{L, R, T\}$
- location of truck: domain  $\{L, R\}$

Aarkov Decision Proces:

Stochastic Shortest Path Problem

Summary 00

# SSP Example



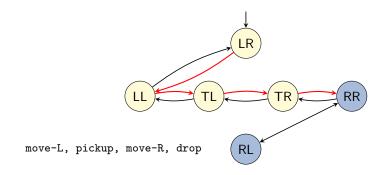
Logistics problem with one package, one truck, two locations:

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

Markov Decision Process 00000000000 Stochastic Shortest Path Problem

Summary 00

# Solutions in Transition Systems



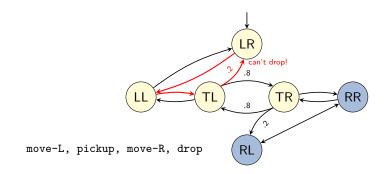
- in a deterministic transition system a solution is a plan, i.e., a sequence of operators that leads from  $s_0$  to some  $s_* \in S_*$
- an optimal solution is a cheapest possible plan
- a deterministic agent that executes a plan will reach the goal

Markov Decision Process

Stochastic Shortest Path Problem

Summary 00

# Solutions in SSPs

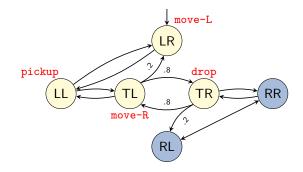


- the same plan does not always work for the probabilistic agent (not reaching the goal or not being able to execute the plan)
- non-determinism can lead to a different outcome than anticipated in the plan
- need again a policy

Markov Decision Proces 00000000000 Stochastic Shortest Path Problem

Summary 00

## Solutions in SSPs



# Policy for SSPs

#### Definition (Policy for SSPs)

Let  $\mathcal{T} = \langle S, A, c, T, s_0, S_* \rangle$  be an SSP. Let  $\pi$  be a mapping  $\pi : S \to A \cup \{\bot\}$  such that  $\pi(s) \in A(s) \cup \{\bot\}$  for all  $s \in S$ .

The set of reachable states  $S_{\pi}(s)$  from s under  $\pi$  is defined recursively as the smallest set satisfying the rules

• 
$$s\in S_{\pi}(s)$$
 and

•  $\operatorname{succ}(s', \pi(s')) \subseteq S_{\pi}(s)$  for all  $s' \in S_{\pi}(s) \setminus S_{\star}$  where  $\pi(s') \neq \bot$ . If  $\pi(s') \neq \bot$  for all  $s' \in S_{\pi}(s_0) \setminus S_{\star}$ , then  $\pi$  is a policy for  $\mathcal{T}$ . If the probability to eventually reach a goal is 1 for all  $s' \in S_{\pi}(s_0)$  then  $\pi$  is a proper policy for  $\mathcal{T}$ .

# Additional Requirements for SSPs

#### We make two requirements for SSPs:

- There is a proper policy.
- Every improper policy incurs infinite cost from every reachable state from which it does not reach a goal with probability 1.
- We will only consider SSPs that satisfy these requirements.
- What does this mean in practise?
  - no unavoidable dead ends
  - no cost-free cyclic behaviour possible
- With these requirements every cost-minimizing policy is a proper policy.

Stochastic Shortest Path Problem

Summary •0

# Summary

# Summary

- There are many planning scenarios beyond classical planning.
- For the rest of the course we consider probabilistic planning.
- (Discounted-reward) MDPs allow state-dependent rewards that are discounted over an infinite horizon
- SSPs are transition systems with a probabilistic transition relation.
- Solutions of SSPs and MDPs are policies.
- For MDPs we want to maximize the expected reward, for SSPs we want to minimize the expected cost.