Planning and Optimization F1. Markov Decision Processes

Gabriele Röger and Thomas Keller

Universität Basel

November 21, 2018

Content of this Course



Motivation

E Search	Q 🕅	Follows to Acid	vity	Test	Time Mart Tax	Har The Fri Sat San Mon San	l 🗠 i				
GMT	10 11 12 12	13 14 15	1	17	1. 19	k9	lest se				
ISS GND SITE	- ALL VHF	ALL VHFALL VHF	ALL VHP		- ALL VH	- ALL VHF					
TDRS ALL							_				
Day Night											
ISS CDR Anton	«свесож- э-сдемит судн-вп-2-с/о судн-овс- прма-с/о судн	-SWTB-C/O MIDDAY-MEAL CKP PEO- MID-3-4-	EVENING PREP-WORK	Q-2 EXERCISE-VELO	PR 36-	PRESLEEP COL	SLEEP-R				
FE-1 Oleg	ADUERCTEX.44. ISS. TEX TEX.44. ISS. CTTC	УУКС- MIDDAY-MEAL IKM-ИЛЛ-INSI	P	EVENING EXERCISE-6,0-2 PREP-WORK	EV OBT OPC	PRESLEEP	SLEEP-R				
FE-2 Drew	A VEN	HM MIDDAY-MEAL JFM F HM JFM-	AC RGN EXERCISE-CEVIS	EXERCISE-ARED	EV SP 081 OP	PRESLEEP	SLEEP-II				
FE-3 Ricky	DRGN-R BAR- APPROACH DRGN- EXERCISE-ARED	MIDDAY-MEAL SPHERES- TEST-S/U SPHERES-1	rslosh-Run		SPHE 081 OPC	PRESLEEP	SLEEP-II				
FE-S Maker	APPROACH DRGN- PRES	L CS- HM CN EXERCISE-ARED	RV DRDRGN/N2-VEST- TELCOPRESS	DRGN-VEST- NOD2- OUTFIT CPA-	EVENING OP	PRESLEEP	SLEEP-II				
FE-6 Nemo	APPROACH DRGN-	MIDDAY-MEAL LTE- L STP	HS- EXERCISE-ARED	DRGN-VEST- NOD2- OUTFIT CPA-	VENING OP	PRESLEEP	SLEEP-D				
CREW	■JPM/CUP-SHUTTER-CLSD	WHC-UNAVAILABLE	WH				HANDS SSC22				
	LAB-SHUTTER-CLOSED						NO T2 E				
	NO NO T2 EXERCISE	NO EXERCISE	-								
COL	COL MPCC ASTROPS-CMD				WISI	ENET-					
	ESA-HRF1-	DMS-HK2 TM PKT-CNFG			_						
		WISE ESA-									
	WISENET-TSLOSH-RUN										
MCC	GO/NO GO DEPT-253 CERSN CAPTURE-GO										

timetable for astronauts on ISS

Generalization of Classical Planning: Temporal Planning

E Search		<u>م</u> 🖄							talam to Asthety						faul Tre	•	1	e Martin Thui fe	Set Sun Mun	i 🗇 🛛
GMT	11.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1	L., W. L.	4	ال ال	¹³ .1.	durte	14.1		1	dad	1	u hu tu	Rational	11 18.111		. 19	hu	1	ulutu R.	
ISS GND SITE	- ALL VHF		LL VHP	-	- ALL VHP			- ALL	VHP	-	ALL VI	9	ALL \	/HP		A1	L VH	IF	- ALL VHP	
TORS ALL		_	_					_	_				_	_			-		_	
Day Night				_				_									2			_
ISS CDR Anton	≪свесож- 5-едемит	-8R-2-C/O	СУДН-ОВС- ПУМА-С/О	судн-	smis-c/o	MIDDAY	MEAL	CKP MD-	PEO- 3-4-	EVER PREI	NING P-WORK	DRRGSE-	6,Q-2	DERCISE-V	tto	EV MB PR 36-	0P	PRESLEEP		COLSLEEP-R EM
FE-1 Oleg	AREA PREP	55. TI REW- 4	TEX-44- ISS- CLSOUT CREW-		PREP	MIDDAY	MEAL	rk M	илл-эмэр				EVENING PREP-WORK	EXERCISE-6,	Q-2	EV OB	OP	PRESLEEP		SLEEP-E
FE-2 Drew	e BS-CHEW-	HMS- WDR5CA		VE V	HM MIDDA	Y-MEAL	JFM WH		HM JFM. PM WHC	RG	EXE C	RCISE-CEVIS	5 EXERCISE-AR	IED	E) Pi	SP 08	OP	PRESLEEP		SLEEP-I
FE-3 Ricky	DRGN-R BAR APPROACH	SSRMS- DRGN-	EXERCISE-ARED		MIDDAY-P	MEAL	SPHERES TEST-S/I	s. U	SPHERES-TSL	.OSH-R	UN .				SP TE	HE OB ST. DRI	OP	PRESLEEP		SLEEP-D
FE-S Maker	ADRGN-R BAR- APPROACH	SSRMS- DRGN-	PRES IC		L (S. 003	HN		EXERC	ISE-ARED	RV	DE DRGN. COPRESS	N2-VEST-	DRGN-VEST- OUTFIT	N002- CPA-	E) 91	ENING IEP-	0Þ	PRESLEEP		SLEEP-D
FE-6 Nemo	4DRGN-R BAR- APPROACH	SSRMS- DRGN-	EXERCISE-CEVIS	LT M	MIDDAY-M	LAL	LTE- METER		STPHS- KE-	- Ex	ERCISE-AR	ED	DRGN-VEST- OUTFIT	N002- CPA-	EVE PRE	NING P.	0P	PRESLEEP		SLEEP-D
CREW	JPM/CUP-SHUTTER-CLSD						WHC-UNAVAILABLE WHH													HANDS SSC22
	LAB-SHUTTER-CLO																NO T2 E			
		NO EXERCISE	NO T2 EXERCISE		NO EXI	ERCISE														
COL	1	COL M	PCC ASTROPS-CMC														WES DE/	IENET-		
	ESA-HRF1- ACT							DMS-HK2 TM PKT-CNFG												
	W25t								ISA. HITL:											
				WISENET-TSLOSH-RUN																
MCC COORD	GO/NO GO DEPT-250 CORSN-CAPTURE-GO																			

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



kinematics of robotic arm

Summary 00

Generalization of Classical Planning: Numeric Planning



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



satellite takes images of patches on earth

Summary 00

Generalization of Classical Planning: MDPs



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





Generalization of Classical Planning: Multiplayer Games



- Chess
- there is an opponent with
- contradictory objective





Summary 00

Generalization of Classical Planning: POMDPs



Solitaire

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

- 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 (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 $\mathcal{T} = \langle S, L, c, T, s_0, S_\star
angle$ 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
- $\mathcal{T} = \langle S, L, c, T, s_0, S_\star
 angle$, 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, ℓ, s') = 1, we also write s → s' or s → 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 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 $s' \in \operatorname{succ}(s, \ell)$ for some ℓ
- with s' ~ succ(s, ℓ) we denote that successor s' ∈ succ(s, ℓ) 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 s⁰ (p₁:ℓ₁/s¹, ..., sⁿ⁻¹ (p_n:ℓ_n/sⁿ s.t. s⁰ = s and sⁿ = s'
Note: n = 0 possible; then s = s'
s⁰,..., sⁿ is called (state) path from s to s'
ℓ₁,..., ℓ_n is called (label) path from s to s'
s⁰ (ℓ₁/s¹, ..., sⁿ⁻¹ (ℓ_n/s) sⁿ is called trace from s to s'
length of path/trace is n
cost of label path/trace is ∑ⁿ_{i=1} c(ℓ_i)
probability of path/trace is ∏ⁿ_{i=1} 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,
 - $R: S \times L \rightarrow \mathbb{R}$ 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,
- $R: S \times L \rightarrow \mathbb{R}$ 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)
- (4,3) 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