Probabilistic (Markov) planning approaches, Markov Decision Processes (MDP)

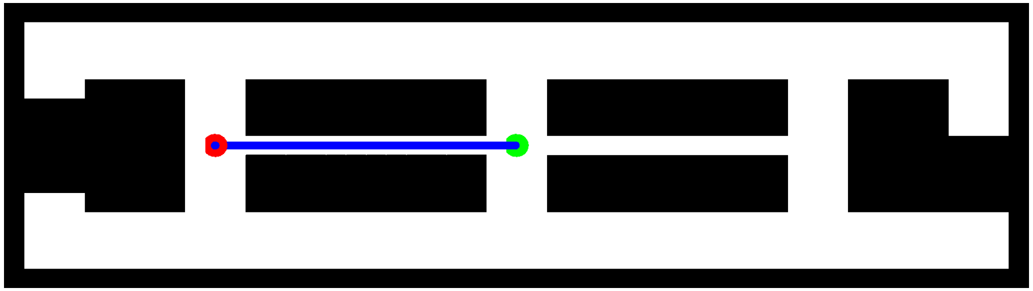
Contents:

- Probabilistic planning – the motivation
- Uncertainty in action selection
  - Markov decision processes
  - Strategy of the planning (the control policy)
  - Partially observable Markov decision processes
- Strategy of the planning – an iterative approach
- The planning goal and its' costs/payoff
  - Objective function construction, planning path payoff/reward
  - Planning horizon choice
  - Cumulative objective/reward function and exponential decay
  - Special cases: Greedy approach, finite horizon, infinite horizon
  - Optimal strategy for fully observable cases, Bellmans' law/equation
- Computing the payoff function
- Applications in robotics
- References

Core problem classes in question?

- Deterministic vs. stochastic actions
- Fully observable vs. partially observable environment

Derministic action, fully observable environment

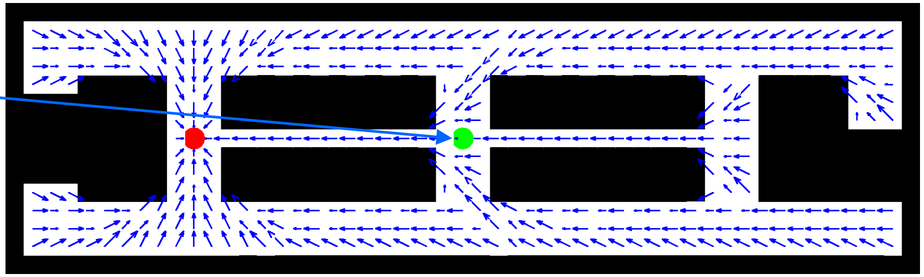


- Imagine the case: The world stands known, „nearly“ symmetric, exhibits narrow and broad passageways; the robot is located in its‘ center (green blob) and not being aware of its‘ orientation (heading) targets towards the goal (red blob). Neither decisions are needed, nor uncertainty is present.
- The task for the robot is to attain the target (red blob)

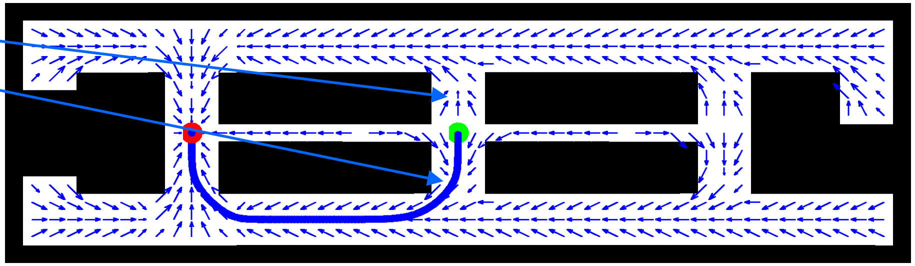
Stochastic action, fully observable environment (Markov Decision Process, MDP)

Payoff function and strategy in MDP:

(a) Deterministic consequence of an applied action. Only unique pathway is possible.



(b) Non-deterministic consequence of an applied action. Multiple pathways are possible.

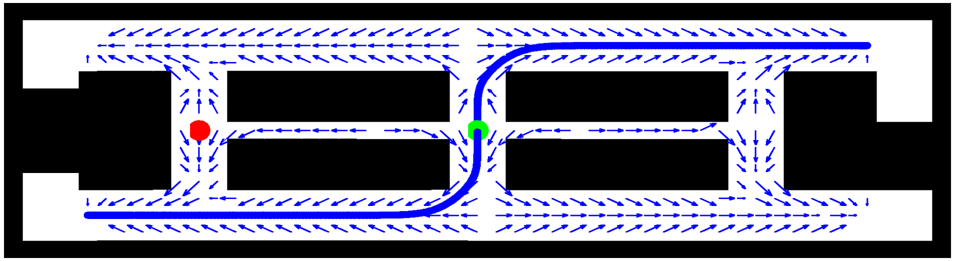


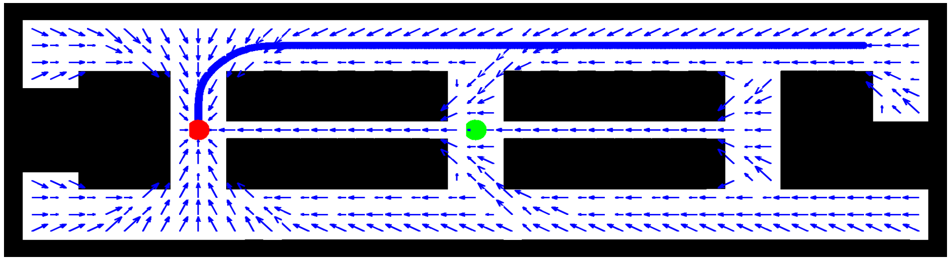
Using deterministic models the robot:

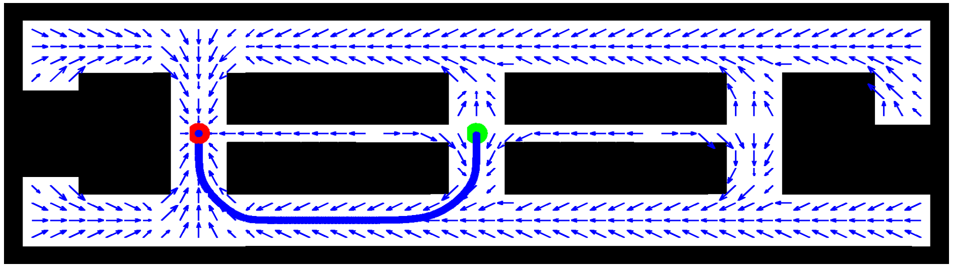
- More easily (reliably) navigates through narrow passageways
- Preferes even longer path to cases, in which action outputs are uncertain. The aim is to supress the risk of a collision.

Stochastic action, partially observable environment (POMDP)

- Actions are „also“ dedicated to knowledge gathering in POMDP cases.
  - i.e. to attain the goal (red blob) with higher certainty than 50%, the probabilistic planner first navigates to a spot, where global orientation can be built.
- Situation (upper case) shows a corresponding strategy and possible alternatives, that the robot/planner can choose to determine its‘ own position.
- Based on knowing its‘ own position, the robot/planner can find initial positions, which allow safe reaching of the goal (red blob)



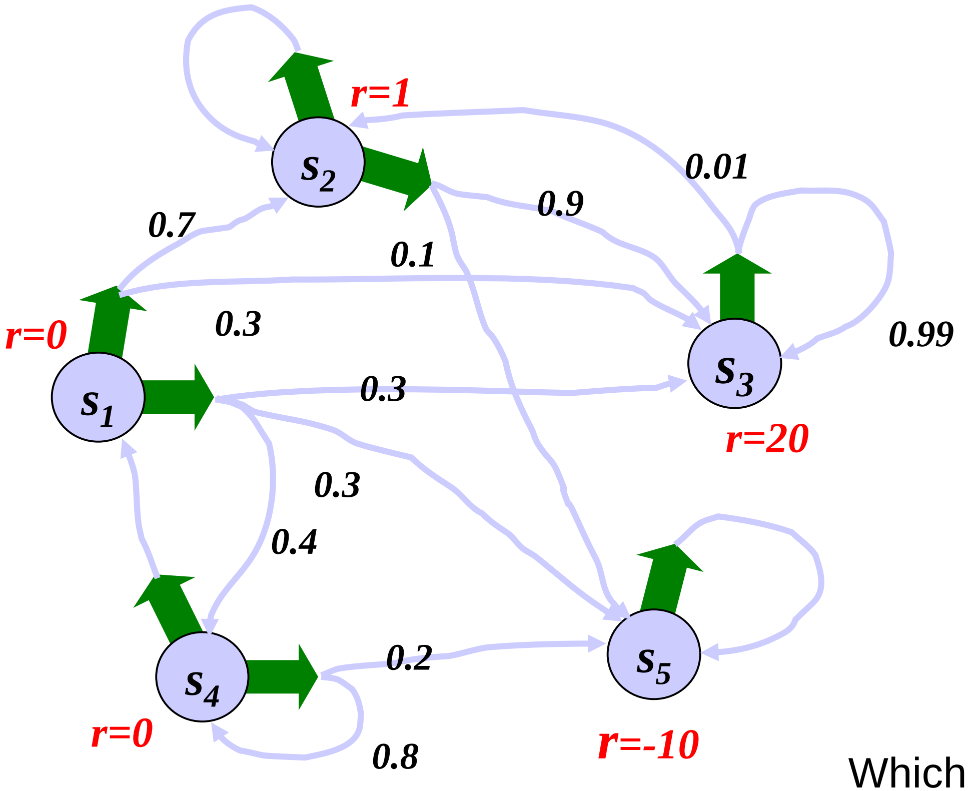




(2 diverse cases are shown in the middle and the bottom figure)

Markov decision process (Markov Decision Process - MDP)

An example of Markov model (a graph) with state(s) ***s***, and probabilities of transitions from ***<0,1>*** and reward(s) ***r*** for reaching a state



Which state is the Goal then?

Markov Decision Process (MDP)

Let‘s presume:

- System states: x
- Admissible actions (controls): u
- Probabilities of transitions: $u,x \rightarrow x'$: $p(x'|u,x)$
- Reward function for reaching a state: $r(x,u)$

The problem statement:

A search for a strategy $\pi(x)$, that maximizes the expected future reward $r(x,u)$.

Reward, Costst and Strategy I

- A strategy (generally said) z_t stands for state observation, that has been achieved via applying of action u_t: $p: \; z_{1:t-1}, u_{1:t-1} \rightarrow \; u_t$
- The Strategy in a fully observable case: $\pi: \; x_t \rightarrow \; u_t$
- Reward for the goal achievement is measured in a quantitative way, and comprises 2 basic and complementary components:

  1. *Cost* (Value) function, that denotes “costs“ of execution of the given path. Measures „price for an action“
  2. Reward (Payoff) function for attaining a state of the system, or a goal. Mesures „success of an action execution“

- Both the preceeding components are integrated into a joint and unique Objective function.
- The Objective function comprises future efforts spent on the next steps via the Cost/Value function and price for the so far executed steps through the Reward/Payoff function.

- In cases of uncertain position of the robot, the approach invokes reasoning in direction as:

„Is raising certainty of reaching the desired goal worth the expected spent efforts?“

Strategy choice I

- Expected (*E - expectation*) cumulative reward with decay *γ*: $R_T = E\left[\sum_{\tau=1}^{T} \gamma^{\tau} r_{t+\tau}\right]$

Types of strategies

- T=1: single-step „greedy“ strategy
- T>1: finite horizon cases, finite reward - typically with no exponential decay, $\gamma = 1$
- T→∞: infinite horizon case, finite reward on condition of exponential decay with $\gamma < 1$ (the series converges for every $r \leq r_{max}$)

- Expected cumulative reward for applying the strategy $R_T^{\pi}(x_t) = E\left[\sum_{\tau=1}^{T} \gamma^{\tau} r_{t+\tau} \,|\, u_{t+\tau} = \pi\,(z_{1:t+\tau-1} u_{1:t+\tau-1})\right]$
- Best possible (optimal) strategy $\pi^* = \underset{\pi}{\operatorname{argmax}} \quad R_T^{\pi}(x_t)$

Strategy alternatives may be:

Single-step strategy:

- Exhibits best possible (optimal) strategy as $\pi_1(x) = \underset{u}{\operatorname{argmax}} \quad r(x,u)$
- The path cost function for single-step strategy as $V_1(x) = \gamma \max_{u} \; r(x,u)$

Strategy choice II

2 – step strategy:

- Optimal strategy $\pi_2(x) = \operatorname*{argmax}_u \left[r(x,u) + \int V_1(x')\, p(x'|u,x)dx' \right]$
- Path cost function $V_2(x) = \gamma \max_u \left[r(x,u) + \int V_1(x')\, p(x'|u,x)dx' \right]$

T – step strategy, or infinite horizon strategy:

- Optimal strategy $\pi_T(x) = \operatorname*{argmax}_u \left[r(x,u) + \int V_{T-1}(x')\, p(x'|u,x)dx' \right]$
- Path cost function $V_T(x) = \gamma \max_u \left[r(x,u) + \int V_{T-1}(x')\, p(x'|u,x)dx' \right]$

...or eventually the infinite horizon case as: $V_\infty(x) = \gamma \max_u \left[r(x,u) + \int V_\infty(x')\, p(x'|u,x)dx' \right]$

which for T→ ∞ exhibits steady value of $V_\infty(x)$ being entitled as *Bellman law/equation*.

Lemma: Every value $V(x)$, that satisfies the Bellman law stands for both the necessary and satisfying conditions of optimality of teh corresponding strategy.

The strategy and costs iteration

- An algorithm to attain (iterate) optimal costs of the path in infinite state space

(For spaces with finite number of states the integration can be replaced by summing over these states):

```
for all x do    {iniciation of the V(x) value}
$\hat{V}(x) \leftarrow r_{\min}$
endfor
```

Eventually in a discrete form: $\hat{V}(x_i) \leftarrow r_{\min}$

```
repeat until convergence
    for all x do
        $\hat{V}(x) \leftarrow \gamma \max_u \left[ r(x,u) + \int \hat{V}(x') p(x'|u,x) dx' \right]$
    endfor
endrepeat
```

Eventually in a discrete form in finite state-spaces:

$$\hat{V}(x_i) \leftarrow \gamma \max_u \left[r(x_i,u) + \sum_{j=1}^{N} \hat{V}(x_j) p(x_j \,|\, u, x_i) \right]$$

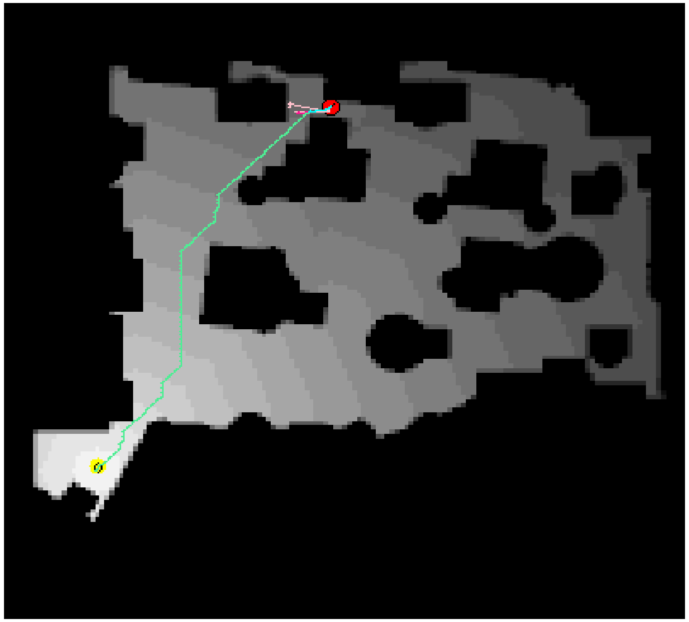
while the optimal strategy (an iteration of the strategy) $MDP(x,\hat{V}) = \pi(x)$ can be calculated from the expression:

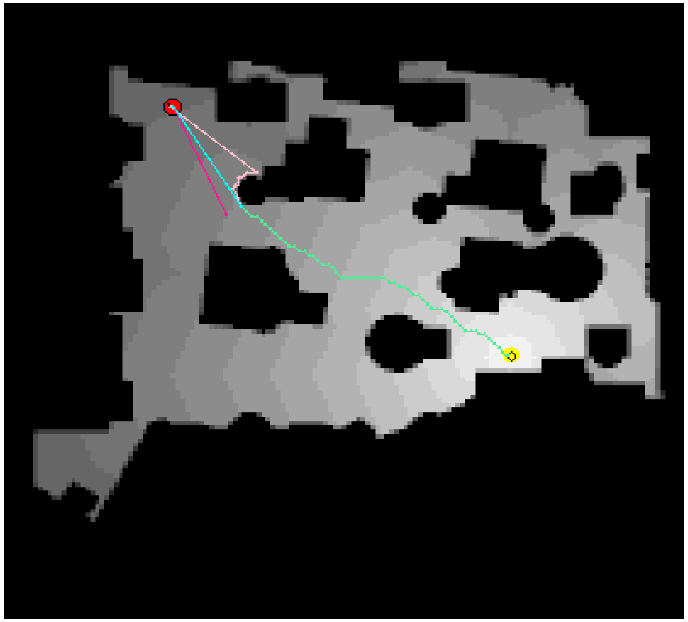
$$\pi(x) = \underset{u}{\operatorname{argmax}} \left[r(x,u) + \int \hat{V}(x') p(x'|u,x) dx' \right]$$

…or in a dicrete form:

$$\pi(x) = \arg \max_u \left[r(x,u) + \sum_{j=1}^{N} \hat{V}(x_j) p(x_j \,|\, u, x_i) \right]$$

An example – robot motion planning





- Obstacles (black spots), the graytone area stands for the cost function *V(x)* – higher values correspond to lighter spots. Applying the „greedy“ strategy using the given cost function leads always to a solution (on condition, that position of the robot is observable)
- Important point is, that the cost function is defined for the whole environment, that allows to determine a strategy even in the case, in which the robot position is not precisely known (uncertain)
- What represents the action(s) in this case?
- How would look a case with non-deterministic actions?

Costs iteration and/or the strategy choice? The experience...

- The best possible/optimal strategy is typically achieved before the price of the path konverges (gets settled).

A varying cost function?

- Iteration of a strategy calculates/determines yet a new strategy. The new strategy is always based on the present cost function. The new strategy subsequently denotes a new cost function.
- The previous process mainly konverges to optimal much faster than in the cases which fix the cost function.

References:

- Thrun S., Burgard W., Fox D.: *Probabilistic Robotics*, The MIT Press, Cambridge, Massachusetts, London, England, 2005, 647 pp., ISBN 0-262-20162-3 (Chapter 14, p.487-p.511)
- http://cs.wikipedia.org/wiki/Markov%C5%AFv_rozhodovac%C3%AD_proces