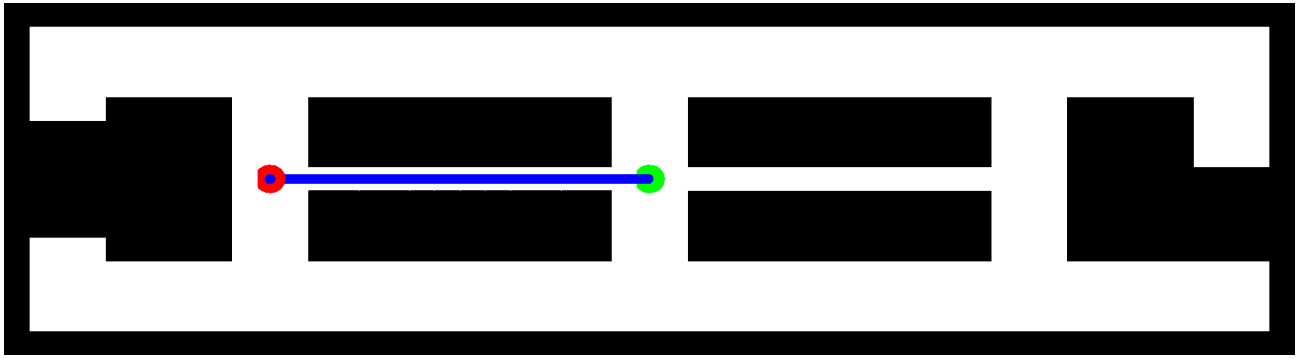
Pravděpodobnostní (Markovské) metody plánování, MDP - obsah

- Pravděpodobnostní plánování - motivace.
- Nejistota ve výběru akce
  - Markovské rozhodovací procesy
  - Strategie plánu (control policy)
  - Částečně pozorovatelné Markovské rozhodovací procesy
- Strategie plánu - metoda iterace
- Cíl a cena za jeho dosažení (payoff/reward)
  - Konstrukce funkce ceny cesty a odměny
  - Plánovací horizont
  - Kumulativní funkce odměny a exponenciální zapomínání
  - Greedy situace, konečný horizont, nekonečný horizont
- Optimální strategie pro plně pozorovatelný případ, Bellmanova rce.
- Výpočet ceny funkce
- Užití v robotice
- Reference

Třídy problémů

- Deterministické vs. stochastické akce
- Plně vs. částečně pozorovatelné prostředí

Deterministické, plně pozorovatelné



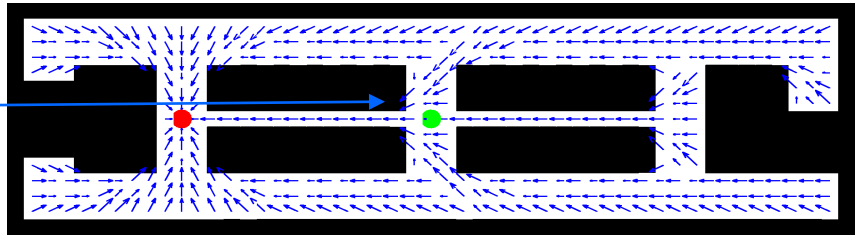
Prostředí je „téměř“ symetrické s úzkými a širokými průchody, robot se nachází ve středu (zelený bod) bez znalosti své orientace a míří do cíle (červený bod). Úkolem robotu je dosáhnout (červeného) cíle.

Stochastické, plně pozorovatelné (Markov Decision Process, MDP)

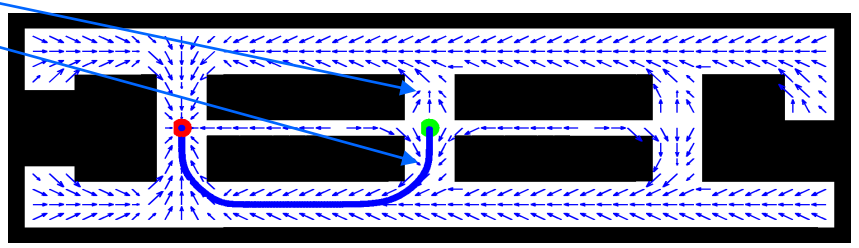
Cenová funkce a strategie v MDP:

(a) Deterministický důsledek akce

(b) Nedeterministický důsledek aplikované akce – umožňuje více cest



V deterministickém modelu robot snadno naviguje úzkými koridory a preferuje delší cestu v případě, že výstupy akce(akcí) jsou nejisté za účelem snížení rizika kolize



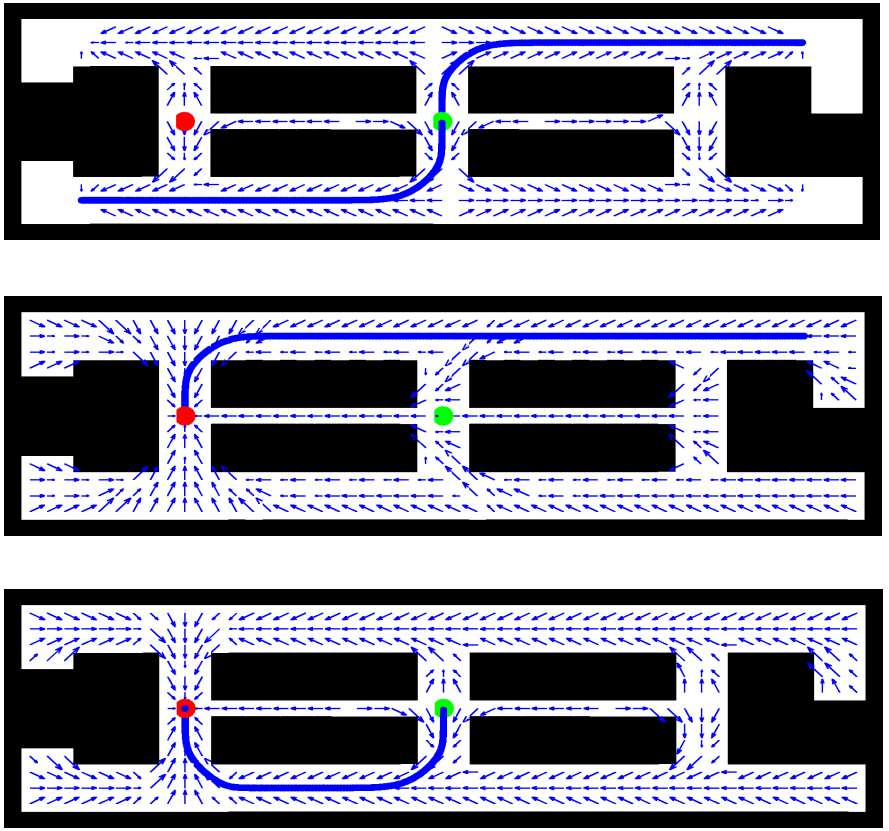
Stochastické, částečně pozorovatelné (Partially Observable MDP, POMDP)

Akce k získávání znalostí v POMDP:

K dosažení cíle (červený bod) s jistotou větší než 50%, plánovač pracující s věrohodností nejprve naviguje do místa, kde může být stanovena globální orientace.

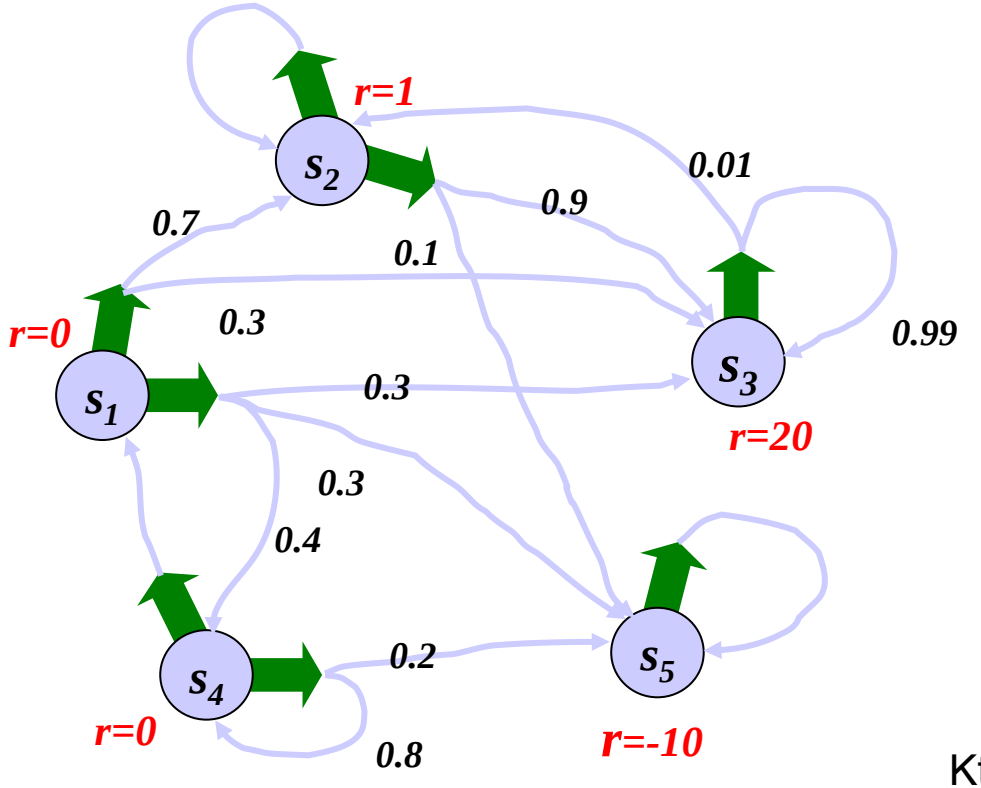
(a) Situace (nahoře) ukazuje odpovídající strategii a možné cesty, jenž může robot zvolit.

(b) V závislosti na znalosti vlastní pozice, robot v prostředí (b) nebo (c) (střed a dole) může stanovit, odkud lze bezpečně dosáhnout cíle.



Markovský rozhodovací proces (Markov Decision Process - MDP)

Příklad Markovského modelu (grafu) se stavem *s*, pravděpodobností přechou *<0,1>* a odměnou za dosažení stavu *r*



Který stav je cílový?

Markovský rozhodovací proces (MDP)

Zadání:

- Stavy systému: x
- Přípustné akce: u
- Pravděpodobnosti přechodů $u,x \rightarrow x'$: $p(x'|u,x)$
- Funkce odměny (reward) za dosažení stavu: $r(x,u)$

Úloha - hledáme:

- Strategii $p(x)$, jenž maximalizuje budoucí očekávanou odměnu $r(x,u)$

Odměny a strategie I

- Strategie (obecný případ), z_t značí pozorování stavu dosaženého akcí u_t:

$$\pi: \quad z_{1:t-1}, u_{1:t-1} \rightarrow \quad u_t$$

- Strategie (plně pozorovatelný případ): $\pi: \quad x_t \rightarrow \quad u_t$

- Cíl a odměna za jeho dosažení – je kvantitativně hodnocena, skládá se ze dvou komplementárních komponent:
  1. *Ceny* (Value function) vyjadřující "náklady" na realizaci dané cesty, měří cenu za akci.
  2. *Odměny* (Reward, Payoff) za dosažení stavu/cíle, měří úspěšnost akce.

Obě předchozí kritéria se integrují do společné cenové funkce (Payoff function) jenž postihuje jednak cenu dosud vykonané cesty a jednak odměnu za dosažený stav, popř. cíl. Takové řešení umožňuje uvažovat i v situacích, kdy robot má nejistou polozici a musí uvažovat způsobem: *„Stojí zvyšující se pravděpodobnost dosažení požadovaného cíle za vynaložené úsilí?"*

Volba strategie I

- Očekávaná (*E - expectation*) kumulativní odměna se zapomínáním γ: $R_T = E\left[\sum_{\tau=1}^{T} \gamma^{\tau} r_{t+\tau}\right]$

Typy strategií:

- T=1: „greedy“ strategie
- T>1: situace s konečným horizontem, typicky bez exp. zapomínání, $\gamma = 1$
- T→∞: situace s nekonečným horizontem, konečná odměna za podmínky exp. zapomínání je s koeficientem $\gamma < 1$ (řada konverguje, pro každé $r \le r_{max}$)

- Očekávaná kumulativní odměna za strategii: $R_T^{\pi}(x_t) = E\left[\sum_{\tau=1}^{T} \gamma^{\tau} r_{t+\tau} \,|\, u_{t+\tau} = \pi(z_{1:t+\tau-1} u_{1:t+\tau-1})\right]$
- Optimální strategie: $\pi^* = \underset{\pi}{\operatorname{argmax}} \; R_T^{\pi}(x_t)$

Varianty strategií mohou být:

1-kroková strategie:

- Optimální strategie: $\pi_1(x) = \operatorname{argmax} \; r(x,u)$
- Funkce ceny cesty pro 1-krokovou optimalní strategii: $V_1(x) = \gamma \max_u \; r(x,u)$

Volba strategie II

2 - kroková strategie:

- Optimální strategie: $\pi_2(x) = \underset{u}{\operatorname{argmax}} \left[r(x,u) + \int V_1(x') p(x'|u,x) dx' \right]$
- Funkce ceny: $V_2(x) = \gamma \max_u \left[r(x,u) + \int V_1(x') p(x'|u,x) dx' \right]$

T - kroková strategie a popř. nekonečný horizont:

- Optimální strategie: $\pi_T(x) = \underset{u}{\operatorname{argmax}} \left[r(x,u) + \int V_{T-1}(x') p(x'|u,x) dx' \right]$
- Funkce ceny: $V_T(x) = \gamma \max_u \left[r(x,u) + \int V_{T-1}(x') p(x'|u,x) dx' \right]$

popř. : $V_\infty(x) = \gamma \max_u \left[r(x,u) + \int V_\infty(x') p(x'|u,x) dx' \right]$ jenž pro T$\rightarrow \infty$ vede k ustálené hodnotě $V_\infty(x)$ a je označována jako *Bellmanova rce*.

Lemma: Každá hodnota $V(x)$ splňující *Bellmanovu rci* je *nutnou* i ***postačující*** podmínkou optimality odpovídající strategie.

Iterace ceny a strategie

Algoritmus k dosažení (iteraci) optimální ceny cesty v nekonečném stavovém prostoru (pro prostory s konečným počtem stavů, lze integrál nahradit součtem přes stavy):

for all x do *{inicializace hodnot V(x)}*
$\hat{V}(x) \leftarrow r_{\min}$
endfor

popř. v diskr. podobě: $\hat{V}(x_i) \leftarrow r_{\min}$

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

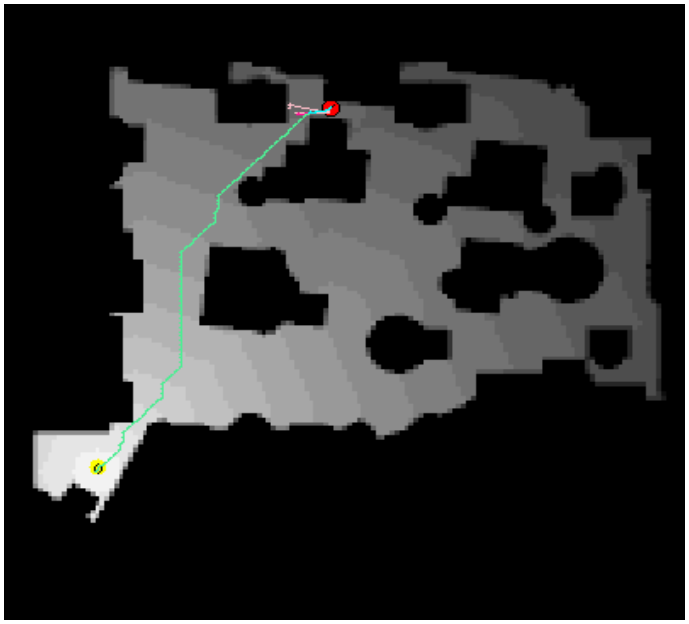
popř. v diskr. podobě pro konečné stavové prostory:
$\hat{V}(x_i) \leftarrow \gamma \max_u \left[r(x_i,u) + \sum_{j=1}^{N} \hat{V}(x_j) p(x_j|u,x_i) \right]$

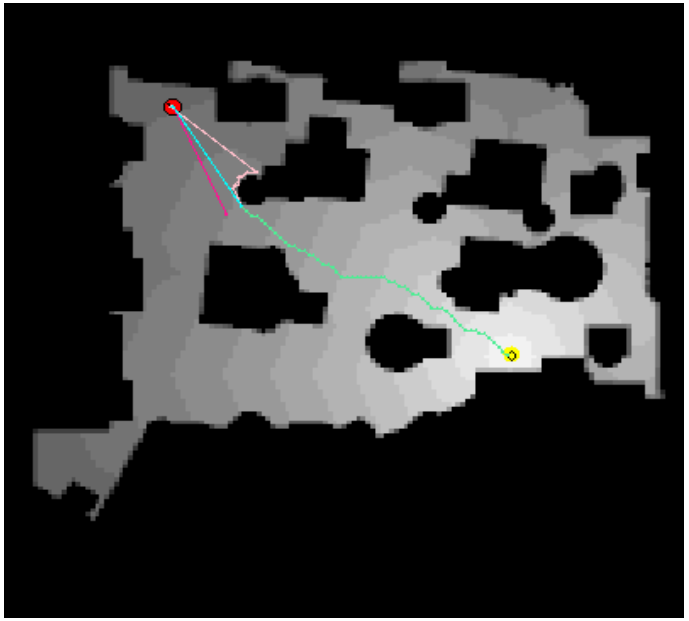
Přičemž optimální strategii (iteraci strategie) $MDP(x,\hat{V}) = \pi(x)$ lze určit prostým výpočtem ze vztahu:

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

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

Příklad - plánování pohybu robotu





- Překážky (černá), cenová funkce *V(x)* je vyjádřena šedou oblastí (vyšší hodnota odpovídá světlejší šedi). „Hladová“ strategie podle hodnot cenové funkce vede k řešení (za předpokladu, že pozice robotu je pozorovatelná)
- Důležitou vlastností je, že cenová funkce je definována pro celé prostředí, což umožní nalézt strategii i v případě, kdy pozice robotu není přesně známa (je nejistá)

Iterace ceny a/nebo strategie ?

- Optimální strategie bývá často dosaženo dříve než dojde ke konvergenci ceny cesty.
- Iterace strategie vypočítává/určuje novou strategii, která je založena na současné cenové funkci. Nově určená strategie následně určí novu cenovou funkci.
- Předchozí proces zhusta konverguje k optimální strategii rychleji.

Reference:

- 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