Akademie věd České republiky
Ústav teorie informace a automatizace, v.v.i.

Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation

RESEARCH REPORT

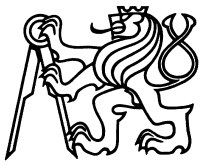
Tereza Siváková

Algoritmický výběr dosažitelných preferencí

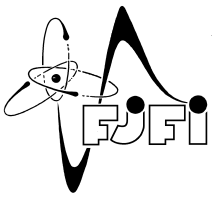
No. 2384 August 19, 2020

ÚTIA AV ČR, P.O.Box 18, 182 08 Prague, Czech Republic
Tel: (+420)266052422, Fax: (+420)286890378, Url: http://www.utia.cas.cz, E-mail: utia@utia.cas.cz

This report presents a draft of a manuscript, which is intended to be submitted for publication. Any opinions and conclusions expressed in this report are those of the authors and do not necessarily represent the views of the involved institutions.



České vysoké učení technické v Praze
Fakulta jaderná a fyzikálně inženýrská



Algoritmický výběr dosažitelných preferencí

Algorithmic Selection of Feasible Preferences

Bakalářská práce

Autor: **Tereza Siváková**

Vedoucí práce: **Ing. Miroslav Kárný, DrSc.**

Akademický rok: 2019/2020

České vysoké učení technické v Praze, Fakulta jaderná a fyzikálně inženýrská

Katedra: matematiky

Akademický rok: 2018/2019

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Student:	Tereza Siváková
Studijní program:	Aplikace přírodních věd
Obor:	Matematické inženýrství
Zaměření:	Aplikované matematicko-stochastické metody
Název práce (česky):	Algoritmický výběr dosažitelných preferencí
Název práce (anglicky):	Algorithmic elicitation of feasible preferences

Pokyny pro vypracování:

1. Seznamte se s teorií markovských rozhodovacích procesů pro případ konečného počtu stavů a akcí.
2. Seznamte se se stavem problematiky kvantitativního popisu cílů rozhodování.
3. Seznamte se s technikou, jak návrh rozhodovací politiky formulovat a řešit jako tzv. plně pravděpodobnostní návrh. V jeho rámci jsou cíle popsány tzv. ideální distribucí.
4. Navrhněte kvalifikaci cílů jako vhodnou optimalizační úlohu na množině možných ideálních distribucí.
5. Výslednou optimalizaci řešte pro jednoduchý případ markovského rozhodovacího procesu s daným modelem okolí a simulačně otestujte.

Doporučená literatura:

1. M. Puterman, Markov decision processes. John Wiley & Sons, NY, 1994.
2. M. Kárný, T. V. Guy, Fully probabilistic control design. Systems & Control Letters, 55:4, 259–265, 2006.
3. M. Kárný et al, Optimized Bayesian Dynamic Advising: Theory and Algorithms. Springer, London, 2006.

Jméno a pracoviště vedoucího bakalářské práce:

Ing. Miroslav Kárný, DrSc
ÚTIA AV ČR, v.v.i., Pod vodárenskou věží 4, 18 208 Praha 8

Jméno a pracoviště konzultanta:

Datum zadání bakalářské práce: 31.10.2018

Datum odevzdání bakalářské práce: 8.7.2019

Doba platnosti zadání je dva roky od data zadání.

V Praze dne 21. listopadu 2018

.....
garant oboru

.....
vedoucí katedry

.....
děkan

Poděkování:
Chtěla bych zde poděkovat především svému školiteli Ing. Miroslavu Kárnému, DrSc. za pečlivost, profesionalitu, ochotu, vstřícnost a odborný i lidský přístup při vedení mé bakalářské práce. Tato práce byla podpořena MŠMT ČR v rámci projektu LTC18075.

Čestné prohlášení:
Prohlašuji, že jsem tuto práci vypracovala samostatně a uvedla jsem veškerou použitou literaturu.

V Praze dne 7. ledna 2020 Tereza Siváková

Název práce:

Algoritmický výběr dosažitelných preferencí

Autor: Tereza Siváková

Obor: Matematické inženýrství

Zaměření: Aplikované matematicko-stochastické metody

Druh práce: Bakalářská práce

Vedoucí práce: Ing. Miroslav Kárný, DrSc., ÚTIA AV ČR

Abstrakt: Tato bakalářská práce se zabývá teorií optimálního rozhodování pro diskrétní markovský rozhodovací proces z hlediska volby preferencí. Za pomoci plně pravděpodobnostního návrhu, který zavádí tzv. *ideální distribuci chování*, která přiřazuje vysoké hodnoty pravděpodobnosti preferovanému chování a malé hodnoty pravděpodobnosti nežádoucímu chování, se hledá optimální rozhodovací politika. Tato práce obsahuje návod k nalezení optimální ideální distribuce chování a přináší obecnější řešení než řešení dosud známá. Dále přidává možnost respektování další preference, a to na volbu akcí. Vlastnosti výsledného rozhodování jsou ilustrovány simulačními experimenty.

Klíčová slova: rozhodování, pravděpodobnostní návrh politik, kvantifikace cílů

Title:

Algorithmic Selection of Feasible Preferences

Author: Tereza Siváková

Abstract: This bachelor's thesis studies the optimal decision making for a discrete Markov decision process with a focus on preferences. By using a fully probabilistic design that introduces the so-called *ideal behavior distribution*, which has high probability values of preferred behaviors and small probability values of inappropriate behaviors, an optimal decision policy has been found. The thesis constructs an algorithm for selecting the optimal ideal behavior distribution and provides a more general solution than published ones. The thesis also opens a possibility to specify further preferences on selected actions. Properties of the resulting decision making are illustrated on simulated examples.

Key words: decision-making, probabilistic policies, quantification of aims

Obsah

Úvod

Tato bakalářská práce se zabývá rozhodovacími procesy, jež jsou základem každé lidské činnosti. Každý člověk za den provede několik set rozhodnutí. Jak se ale rozhodnout správně a co nejsnadněji dosáhnout vytyčeného cíle?

Na tuto otázku se stále nedá jednoznačně odpovědět. Většina lidských rozhodnutí je citově zabarvena, proto je velice obtížné matematicky popsat chování agenta, který rozhodnutí činí, a dostat tak nejlepší sadu rozhodnutí k dosažení vytyčeného cíle. Také složitost a neurčitost řešeného problému značně komplikuje jednoznačnost odpovědi. Moderní teorie rozhodování se však snaží tuto zátěž z agenta sejmout a získat tak soubor matematických rovnic, jejichž řešením dostane optimální rozhodovací strategii zvanou politika. Když se na problém pohlíží čistě matematicky, dá se pravděpodobnostně popsat nejlepší cesta k vytyčenému cíli. V této práci se využívá markovský rozhodovací proces, který popisuje systém pomocí stavů, ve kterých se nachází, a akcí, které jsou na systému vykonány. Akce, které jsou na systému vykonány jsou vybírány agentem, který svými rozhodnutími ovlivňuje uzavřenou rozhodovací smyčku. Uzavřenou rozhodovací smyčkou se nazývá propojení agenta s náhodně reagujícím systémem.

Markovské rozhodovací procesy jsou procesy, při nichž zaměří-li se agent na určitý stav v čase a zvolí-li na základě pozorovaného stavu akci, vyvolá dva důsledky. Za prvé se systém posune do dalšího stavu dle příslušného rozdělení pravděpodobnosti stavů systému a za druhé obdrží okamžitou ztrátu. Termín "okamžitá ztráta" lze vnímat jako výdej energie k vykonání akce a k posunu do dalšího stavu nebo jako cenu za vychýlení od žádaného stavu. Přirozeně je pro agenta nejlepší, když je tato vydaná energie nebo cena malá. Poté se systém nachází v dalším stavu a agent má před sebou opět volbu akce. Opakováním stejného postupu se získá markovský řetězec, který se dá pomocí určitých matematických operací optimalizovat. Důležitou vlastností markovského rozhodovacího procesu je jeho nezávislost na posloupnosti stavů a akcích minulých. Závisí tedy pouze na pozorovaném stavu a akci, která je vykonaná právě v čase pozorování.

To, do jakého stavu se systém posune, je dáno rozdělením pravděpodobnosti stavů systému. Toto rozdělení je určeno systémem a z části i okolním světem, ve kterém se systém vyskytuje.

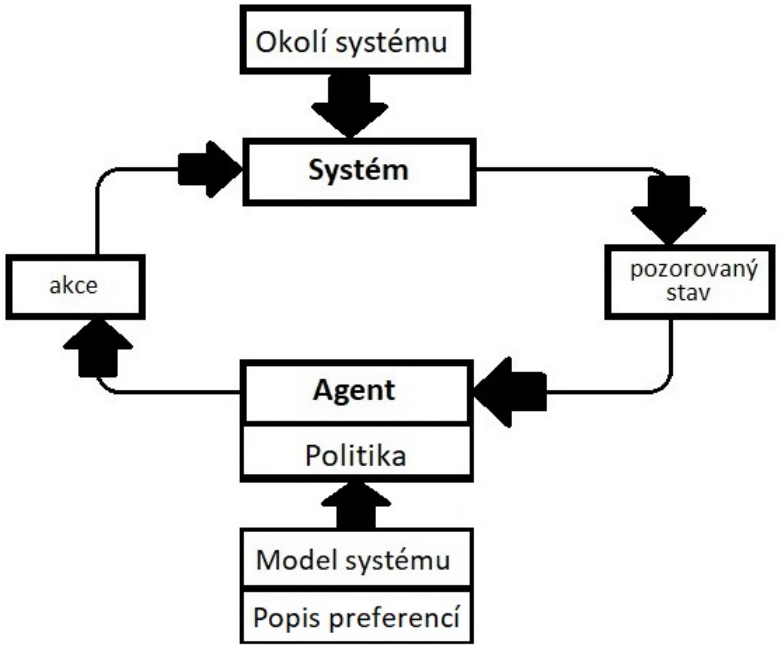
Příklad Agent chce mít v dané místnosti teplotu $25°C$. Daná místnost je systémem, stavy jsou jednotlivé teploty, akce je otáčení kolečka topení doprava/doleva a okolím světem je reálné okolí místnosti. Bude-li tedy venku zima, bude potřeba více topit, otáčet kolečkem topení doprava, aby bylo v místnosti dosaženo požadované teploty. Bude-li ale naopak venku teplo, nebude potřeba topit vůbec a preferovaný stav nastane automaticky. Pravděpodobnostní rozdělení stavů systému, je tedy ovlivněno i okolím, ve kterém se systém vyskytuje.

Je také zjevné, že pro různé systémy budou rozdělení pravděpodobnosti různé, a pro každý systém bude volba též akce dávat různé výsledky.

Rozhodovací politika udávající výběr akce je další klíčový prvek markovského rozhodovacího procesu. Pomocí ní lze získat nejlepší cestu k vytyčenému cíli. Politiku konstruuje agent z modelu systému a popisu preferencí. Model systému, který agentovi dává částečnou znalost systému, vstupuje do uza-

vřené smyčky zvnějšku. Může být zadaný pevně nebo se využije bayesovského učení, které odhaduje model systému v každém čase uzavřené smyčky. Model systému je pak v každém dalším čase pomocí bayesovského učení vylepšován až do takové míry, že má agent téměř úplnou znalost systému. Díky této znalosti je pak schopen nejlépe vybrat akci, která ho posune do preferovaného stavu. Preferovaný stav musí být jednoznačně určen a zadán zvnějšku do uzavřené smyčky. Označuje se jako "popis preferencí".

Tato uzavřena smyčka je vykreslena na obrázku (Obrázek 1). Stěžejním úkolem této práce je nalezení



Obrázek 1: Schéma uzavřené smyčky

optimálního kvantitativního popisu neúplně zadaných preferencí a k němu nalezení optimální politiky, která bude minimalizovat ztráty.

Tato práce konkrétně k rozhodovaní využívá zobecnění markovského rozhodovacího procesu zvané plně pravděpodobnostní návrh [8]. Tento návrh je výpočetně jednodušší a jeho další výhodou je, že k zadávání preferencí využívá tzv. *ideální distribuci chování*, ke které se pak snaží tu reálnou přiblížit. Proto v plně pravděpodobnostním návrhu do uzavřené smyčky jako "popis preferencí" vstupuje ideální distribuce. Ideální distribuce chování má vysoké hodnoty pravděpodobnosti výskytu v žádoucím stavu a malé hodnoty pravděpodobnosti ve stavech nežádoucích. Rozdíl mezi těmito dvěma distribucemi (ideální a reálnou) se pomocí jednoduchých matematických operací minimalizuje, a tím se nalezne optimální rozhodovací politika.

Teorii k markovským rozhodovacím procesům tato práce čerpá z [16]. K řešení rozhodovacích markovských procesů se využívá dynamické programování, které je popsané v [1]. Pro zpřesnění modelu systému tato práce dále využívá již zmíněné bayesovské učení, které je popsané v [7] a [15]. Teorie k nalezení optimální politiky v plně pravděpodobnostním návrhu je čerpaná z [10].

Rozhodovací procesy mají opravdu široké využití a teorie rozhodování je již na vysoké úrovni. Stále se ale potýká s problémem jednoznačného vyjádření úplné preference. Jelikož agent nemusí vždy zcela rozumět matematice rozhodování, může formulovat protichůdné preference nebo preference, které nemusí být v daném systému dosažitelné. Pak není možné zcela vyřešit daný problém. Může se však daným preferencím pomocí optimalizace alespoň přiblížit. Dalším problémem je, že při přidání další preference může mít rozhodovací proces několik možných řešení. Proto se využívá zpětné vazby agenta, který rozhoduje, jestli mu výsledek vyhovuje, nebo chce pokračovat v ladění parametrů k dosažení lepších výsledků. Agent se během celého procesu navíc rozhoduje, které z daných preferencí a řešení bude upřednostněno. Tímto procesem se již zabývalo mnoho článků například [3], [2] nebo [4]. Teorie rozhodování a preference se v praxi využívá i v medicíně, tím se zabýval článek [13], kde se podle preferencí pacienta, lékař rozhoduje o možnosti léčby. Dále se teorie rozhodování a preference využívá ve skupinovém rozhodování [14], [5].

Tato práce se však nejvíce opírá o článek [9], kde se rozhodovací proces s agentovými preferencemi řeší pomocí plně pravděpodobnostního návrhu, jehož výhody jsem již zmínila výše. Plně pravděpodobnostní návrh společně s odhadováním modelu systému bayesovským učením sám o sobě optimalizuje celý rozhodovací proces. Zachází s neúplným popisem agentových preferencí lépe, neboť kromě nich respektuje i model systému. To je hlavní přínos oproti článkům zmíněným výše, které k vylepšení výsledků využívají jen zpětnou vazbu agenta.

Popis preferencí je inspirován výpočtem ideálních pravděpodobností z článku [9]. Avšak má práce přináší několik vylepšení této teorie. Prvním vylepšením je přidání preference na akce. Agent může mít například preference malých akcí kvůli vydání menší energie na vykonání akce. Tato preference se však nemusí vždy slučovat s nejčetnějším výskytem v preferovaném stavu, proto se zde využívá zpětné vazby agenta k ladění parametrů pro dosažení nejlepšího výsledku, jako je tomu v článcích zmíněných výše. Druhým vylepšením je výpočet ideálního modelu okolí obecnější metodou. Množina řešení ideálu z [9] může být v diskrétním světě prázdná a navržená obecnější metoda nalezne řešení vždy.

V první kapitole této práce představuji matematické nástroje, které dále používám. Ve druhé kapitole se zaměřuji na hledání optimálního ideálu a navrhuji zde možná vylepšení. Třetí kapitola obsahuje experimentální část, která ověřuje přínos vylepšené teorie z druhé kapitoly. V poslední kapitole jsou shrnuty dosažené výsledky s hlavními přínosy práce a zmíněny směry dalšího výzkumu.

Kapitola 1

Matematické nástroje

V této kapitole je shrnuta teorie, o kterou se práce opírá.

pojem	**značení**
přirozená čísla	$\mathbb{N}$
množina prvků m	$\mathbf{M}$
m náleží do množiny $\mathbf{M}$	$m \in \mathbf{M}$
střední hodnota	E
pravděpodobnost jevu x	$p(x)$
argument maxima	$\text{argmax}_{x \in \mathbf{M}} f(x)$
mohutnost množiny $\mathbf{M}$	$\lvert\mathbf{M}\rvert$
skalární součin	$< \cdot, \cdot >$

Tabulka 1.1: Značení

- Podmíněná hustota pravděpodobnosti $p(x|y)$ veličiny x za podmínky, že nastalo y je definovaná pomocí sdružené $p(x, y)$ a marginální hustoty $p(y)$ vztahem $p(x|y) = \frac{p(x,y)}{p(y)}$ pro $p(y) > 0$.
- Řetězovým pravidlem pro dva náhodné jevy x, y se rozumí rovnost $p(x, y) = p(x|y)p(y)$, což je vlastně jen jiné vyjádření podmíněné pravděpodobnosti.
- Argument maxima je definován jako $\text{argmax}_{x \in \mathbf{M}} f(x) = \{x \mid x \in \mathbf{M} \wedge \forall y \in \mathbf{M} : f(y) \leq f(x)\}$. Obdobně pro argument minima.

Tato práce většinou uvažuje pouze diskrétní hodnoty x, y, pak hustoty pravděpodobnosti splývají s pravděpodobnostmi.

Definice 1. *Jsou-li p, q dvě hustoty pravděpodobnosti nad množinou $\boldsymbol{X}$, pak se jejich blízkost vyjadřuje pomocí **Kullback-Leiblerovy (KL)** divergence [12]*

$$D(p \parallel q) \equiv \sum_{x \in \boldsymbol{X}} p(x) \ln \frac{p(x)}{q(x)}, \tag{1.1}$$

a $D \geq 0$ pro všechna f, q. Rovnost $D(p \parallel q) = 0$ nastává právě tehdy, když $p = q$ skoro všude na $\boldsymbol{X}$.

Věta 1. *Nechť jsou dány hustoty pravděpodobnosti na spočetné množině stavů* $\mathbf{S}$ *tj.* $f(s), g(s) \geq 0$ *pro všechna* $s \in \boldsymbol{S}$ *a* $\sum_{s\in\boldsymbol{S}} f(s) = 1$, $\sum_{s\in\boldsymbol{S}} g(s) = 1$. *Pak minimalizace*

$$\min_{g(s)} \sum_{s\in\boldsymbol{S}} f(s) \ln\left(\frac{1}{g(s)}\right) \tag{1.2}$$

má řešení $g(s) = f(s)$ *na* $\mathbf{S}$..

Důkaz. Minimum se nalezne pomocí vyšetření extrému funkce s vazební podmínkou $\sum_{s\in\mathbf{S}} g(s) = 1$.

$$L = \sum_{s\in\mathbf{S}} f(s) \ln\left(\frac{1}{g(s)}\right) + \lambda \sum_{s\in\mathbf{S}} g(s). \tag{1.3}$$

Derivace L podle $g(s)$ je rovna

$$-\frac{f(s)}{g(s)} + \lambda = 0$$

$$g(s) = \frac{f(s)}{\lambda}. \tag{1.4}$$

Druhá derivace je pak

$$\frac{f(s)}{g^2(s)} > 0.$$

Po sečtení (1.4) a po dosazení podmínky $\sum_{s\in\mathbf{S}} g(s) = 1$ do (1.4)

$$1 = \sum_{s\in\mathbf{S}} g(s) = \frac{\sum_{s\in\mathbf{S}} f(s)}{\lambda},$$

podle předpokladu i $\sum_{s\in\mathbf{S}} f(s) = 1$

$$1 = \frac{\sum_{s\in\mathbf{S}} f(s)}{\lambda} = \frac{1}{\lambda} \Rightarrow \lambda = 1.$$

Po zpětném dosazení $\lambda = 1$ do (1.4)

$$f(s) = g(s),$$

čímž je věta dokázána. □

1.1 Diskrétní markovský rozhodovací proces

Definice 2. *Diskrétní markovský rozhodovací proces (**MRP**) je definován uspořádanou pěticí* $\{\boldsymbol{T},\boldsymbol{S},\boldsymbol{A}, p, z\}$ *a politikou* π*, kde:*

- $\boldsymbol{T}$ *označuje diskrétní konečný soubor rozhodovacích epoch* $\boldsymbol{T} = \{0, 1, 2, \cdots, |\boldsymbol{T}|\}$, $|\boldsymbol{T}| \in \mathbb{N}$
- $\boldsymbol{S}$ *označuje diskrétní konečnou množinu možných stavů systému* $\boldsymbol{S} = \{s_1, s_2, \cdots, s_n\}$, $n \in \mathbb{N}$
- $\boldsymbol{A}$ *označuje diskrétní konečný soubor možných akcí* $\boldsymbol{A} = \{a_1, a_2, \cdots, a_m\}$, $m \in \mathbb{N}$
- p *je markovská pravděpodobnostní funkce přechodu, kde* $p(\tilde{s} \mid a, s) \geq 0$.
  Platí $\sum_{\tilde{s} \in S} p(\tilde{s} \mid a, s) = 1$ *a* $p(\tilde{s} \mid a, s, v) = p(\tilde{s} \mid a, s)$ *pro* $\forall s \in \boldsymbol{S}$, $\forall a \in \boldsymbol{A}$, *kde* $v \in \boldsymbol{V}$ *jsou další minulá pozorování.*
- z *je reálná pravděpodobnostní funkce ztrát* $z = z(\tilde{s}, a, s)$, *kde* $\tilde{s}, s \in \boldsymbol{S}$, $a \in \boldsymbol{A}$.
  Funkce ztrát hodnotí cestu ze stavu s *pomocí akce* a *do stavu* $\tilde{s}$.
- *Politika* π_t *je posloupnost pravděpodobnostní* $\pi(a \mid s)$, *akcí* $a = a_\tau$ *podmíněných stavy* $s = s_{\tau-1}$, *kde* $s \in \boldsymbol{S}$, $a \in \boldsymbol{A}$, $t \leq \tau \leq |\mathbf{T}| = h$, *nazývaných rozhodovací pravidla a splňujících* $\pi(a \mid s) \geq 0$ *a* $\sum_{a \in \boldsymbol{A}} \pi(a \mid s) = 1$ *pro* $\forall a \in \boldsymbol{A}$, $\forall s \in \boldsymbol{S}$.

Rovnost $p(\tilde{s} \mid a, s, v) = p(\tilde{s} \mid a, s)$ vystihuje, že pravděpodobnost výskytu stavu $\tilde{s}$ nezávisí na posloupnosti stavů minulých, ale pouze na pozorovaném stavu a na zvolené akci. Proto je možné dívat se na jednotlivé vývoje do dalších stavů individuálně jen na základě stavu, který je pozorován, a akce, která je pro tento stav zvolena. Pokud markovský rozhodovací proces závisí na čase, jeho pravděpodobnostní funkce přechodu je rovna $p(\tilde{s} \mid s, a) = p(s_t = \tilde{s} \mid a_t = a, s_{t-1} = s)$, kde s_{t-1} je stav předešlý, a_t je akce vykonaná agentem a s_t je nový stav.

Dalším pojmem je očekávaná ztráta, která před užitím akce a ve stavu s číselně ohodnotí "správnost" výběru akce. Očekávaná ztráta je definovaná jako

$$E[z \mid a, s] = \sum_{\tilde{s} \in \mathbf{S}} z(\tilde{s}, a, s) p(\tilde{s} \mid a, s). \tag{1.5}$$

Funkci ztrát lze fyzikálně vnímat jako cenu odchylky stavu $\tilde{s}$ a jeho změny $\tilde{s} - s$ od hodnot žádaných a také jako množství energie, které je potřeba vydat k přesunu do dalšího stavu. Přirozeně je nejvýhodnější, aby odchylky byly co nejmenší a vydaná energie co nejmenší. Je tedy potřeba volit takové akce, které budou minimalizovat očekávanou ztrátu. Volba akce musí však být činěna před pozorováním stavu $\tilde{s}$, takže je potřeba najít politiku, která budu vybírat akce s nejmenší očekávanou ztrátou. K nalezení takové politiky se využívá dynamické programování.

1.2 Dynamické programování

Nalezení optimální politiky je složitý problém, který nelze řešit standardní minimalizací ztrátové funkce, jelikož je celý markovský proces provázaný závislostmi (na stavu minulém a zvolené akci) a navíc je celý proces náhodný, takže důsledky akce a ve stavu s lze popsat nejvýše pravděpodobnostně. Proto se k nalezení optimální politiky využívá složitější matematický aparát, zvaný dynamické programování, které tyto vlastnosti a požadavky respektuje.

Dynamické programování je obecně metodou pro řešení složitého problému jeho rozdělením na menší podproblémy. V případě markovského rozhodovacího procesu se však dynamické programování využívá od konce, především pro zajištění toho, že pro volbu akce je možno užít jen pozorovaný stav nikoliv stav budoucí. Začíná se ve stavu, do kterého se chce agent dostat a poté se postupnými výpočty vrací zpět do stavů předešlých tak, aby celková očekávaná ztráta byla co nejmenší.

Matematicky lze tuto minimalizaci definovat pomocí pojmu funkce hodnoty politiky [1].

Definice 3. *Nechť je dán* $M = (\boldsymbol{T,S,A}, p, z)$ ***MRP*** *s politikou* $\pi_t \in \mathbf{\Pi}_t$*, pak jsou hodnoty funkce politiky* $\varphi_t^{\pi_t} : \boldsymbol{S} \to \mathbb{R}$, $\forall \pi_t \in \mathbf{\Pi}_t \; \forall t \in \mathbf{T}\backslash\{h\}$*, kde* $h = |\boldsymbol{T}| \in \mathbb{N}$ *se nazývá horizont a je to čas pozorování konečného stavu, definované jako*

$$\varphi_t^{\pi_t}(s) = E[\sum_{h \geq \tau > t, \tau \in \mathbf{T}} z(s_\tau, a_\tau, s_{\tau-1}) \mid s_t = s], \tag{1.6}$$

kde $s_\tau, s_{\tau-1} \in S$ *představují stavy systému v* $\tau, \tau - 1$ *časových epochách,* $s \in S$ *je stav, pro který se vypočítává hodnota politiky* π *a* a_t *je zvolená akce.*

Funkce $\varphi_t^{\pi_t}$ je tedy očekávanou hodnotou funkce kumulovaných ztrát stavu $s \in \mathbf{S}$ v čase $t \in \mathbf{T}$ s politikou π_t. Očekávanou ztrátu je nutné zvolit na začátku celého procesu a zadat ji zvenčí do uzavřené smyčky spolu jako "popis preferencí" (Obrázek 1). Pro získání optimální politiky tudíž stačí nalézt pouze optimální hodnoty této funkce a poté pro každý stav zvolit rozhodovací pravidlo, které vybere akci a, která zaručí dosažení minima v rovnici (1.6). Optimální politikou je potom argument minima optimální hodnoty této funkce

$$\pi_t^0 = argmin\, \varphi_t^{\pi_t}(s). \tag{1.7}$$

Věta 2. *Nechť je dán* $M = (\boldsymbol{T,S,A}, p, z)$*,* ***MRP****, pak se optimální hodnota funkce* $\varphi_t^0(s) = \min_{\pi_t \in \mathbf{\Pi}_t} \varphi_t^{\pi_t}(s)$ *vypočítá rekurzivně* $\forall t \in \mathbf{T}\backslash\{h\}$ *pomocí vzorce*

$$\varphi_t^0(s) = \min_{a \in \mathbf{A}} E[z(\tilde{s}, a, s) + \varphi_{t+1}^0(\tilde{s}) \mid a, s], \tag{1.8}$$

s předpokladem, že $\varphi_h^0 = 0$*, což vyplývá z (Definice 3). Optimální politika* π_t^0*, jako argument minima, je soustředěna na minimalizujících argumentech v rovnici (1.8).*

Důkaz tohoto tvrzení je možné nalézt v [17].

1.3 Plně pravděpodobnostní návrh

Jak je již zmíněno výše, standardní řešení **MRP** je založeno na optimalizaci očekávané hodnoty ztrátové funkce. Je základem většiny systematických řešení rozhodovacích úloh. Má však různá omezení (např. výpočetní složitost či obtížnost kombinace dílčích preferencí). Proto se hledala jiná alternativa a byl navržen plně pravděpodobnostní návrh (**PPN**) [6, 11].

PPN je z hlediska této práce výpočetně méně náročný a jeho hlavní výhodou je, že zavádí tzv. *ideální distribuci chování*, která odráží agentovy preference a ke které se snaží reálné distribuce chování přiblížit. Tento návrh se totiž opírá o fakt, že minimalizace ztrátové funkce se dá chápat jako pokus ovlivnit (pravděpodobnostní) distribuci proměnných uzavřené smyčky.

Byla tedy zavedena ideální distribuce chování uzavřené smyčky, která přiřazuje vysoké hodnoty pravděpodobnosti požadovanému a nízké hodnoty pravděpodobnosti nežádoucímu chování. K přiblížení hodnot distribucí ideálního a reálného chování uzavřené smyčky se využívá minimalizace Kullback-Leiblerovy divergence, která je definovaná výše (Definice 1).

Analogicky s **MRP** využívá **PPN** pravděpodobnostní funkci přechodu a politiku. Jak je již řečeno, k nalezení optimální politiky však nevyužívá funkci ztrát, ale minimalizaci **KL** divergence, která přibližuje ideální distribuci chování k reálné distribuci chování. Politika je pak argumentem minima této divergence.

Ideální distribuci je opět nutné agentovi zadat zvenčí stejně jako ztrátovou funkci (Obrázek 1). Stěžejním úkolem této práce je nalezení optimální ideální distribuce chování, která by co nejlépe respektovala neúplně a neformálně popsané preference, následně našla optimální politiku a demonstrovala její vlastnosti experimentálně.

Nyní je potřeba tuto myšlenku popsat více matematicky. Agent volí akce $a_t \in \mathbf{A}, t \in \mathbf{T}$, které ovlivňují posunutí systému ze stavu $s_{t-1} \in \mathbf{S}$ do stavu $s_t \in \mathbf{S}$. Soubor vybraných akcí a stavů do daného horizontu $h \in \mathbb{N}$, popisující chování uzavřené smyčky (agent-prostředí), je definován

$$b = (s_0, a_1, s_1, a_2, s_2, \ldots, a_h, s_h) \in \mathbf{B}. \tag{1.9}$$

Výběr akce a_t je dán rozhodovací politikou agenta $\pi \in \mathbf{\Pi}$ skládající se z posloupností rozhodovacích pravidel

$$\pi = (\pi(a_1 \mid s_0), \pi(a_2 \mid s_1), \ldots, \pi(a_h \mid s_{h-1})). \tag{1.10}$$

Chování uzavřené smyčky je pak plně popsáno pomocí hustoty pravděpodobnosti $c^\pi(b)$ závisející na π. Pomocí řetězového pravidla lze hustotu pravděpodobnosti zapsat takto

$$c^\pi(b) = \prod_{t \in \mathbf{T}} p(s_t \mid a_t, s_{t-1}) \pi(a_t \mid s_{t-1}) p(s_0), \tag{1.11}$$

kde hustota pravděpodobnosti $p(s_t \mid a_t, s_{t-1})$ popisuje dynamiku systému a $\pi(a_t \mid s_{t-1})$ je náhodné rozhodovací pravidlo. Posloupnost hustot pravděpodobností

$$p = (p(s_1 \mid a_1, s_0), p(s_2 \mid a_2, s_1), \ldots, p(s_h \mid a_h, s_{h-1})) \tag{1.12}$$

tvoří známý a pevný model systému. Jak je již zmíněno výše, distribuce chování uzavřené smyčky $c^\pi(b)$ je potřeba co nejblíže přiblížit ideální distribuci chování uzavřené smyčky, dané vztahem

$$c^i(b) = \prod_{t \in \mathbf{T}} p^i(s_t \mid a_t, s_{t-1}) \pi^i(a_t \mid s_{t-1}) p^i(s_0). \tag{1.13}$$

Podobně jako v případě $c^\pi(b)$, je $p^i(s_t \mid a_t, s_{t-1})$ ideální model systému, reprezentující dynamiku ideálního systému, a $\pi^i(a_t \mid s_{t-1})$ je ideální rozhodovací pravidlo. Hodnoty těchto dvou uzavřených smyček lze přiblížit pomocí minimalizace Kullback-Leiblerovy divergence, jak je již zmíněno výše. Optimální hodnota rozhodovací politiky je pak dána

$$\pi^0 \in \operatorname{Arg\,min} D(c^\pi \| c^i) \tag{1.14}$$

$$D(c^\pi \| c^i) = \sum_{b \in \mathbf{B}} c^\pi(b) \ln\Big(\frac{c^\pi(b)}{c^i(b)}\Big).$$

Věta 3. *Rozhodovací pravidla tvořící optimální rozhodovací politiku π^0 se vypočítají pomocí vzorce*

$$\pi^0(a_t \mid s_{t-1}) = \pi^i(a_t \mid s_{t-1}) \frac{\exp[-d(a_t, s_{t-1})]}{\gamma(s_{t-1})}, \tag{1.15}$$

kde

$$d(a_t, s_{t-1}) = \sum_{s_t \in \boldsymbol{S}} p(s_t \mid a_t, s_{t-1}) \ln \left[\frac{p(s_t \mid a_t, s_{t-1})}{\gamma(s_t) p^i(s_t \mid a_t, s_{t-1})} \right]$$
$$\gamma(s_{t-1}) = \sum_{a_t \in \boldsymbol{A}} \pi^i(a_t \mid s_{t-1}) \exp[-d(a_t, s_{t-1})], \tag{1.16}$$

$t = h, h-1, \ldots, 1$.

Zpětná rekurze začíná v $\gamma(s_h) = 1 \geq \gamma(s_t), \forall t \in \boldsymbol{T}$. *Dosažené minimum je*

$$\min_{\pi \in \Pi} D(c^\pi \| c^i) = -\ln(\gamma(s_0)). \tag{1.17}$$

Hodnota funkce $-\ln(\gamma(s_{t-1}))$ *je omezená shora*

$$-\ln(\gamma(s_{t-1})) \leq -\ln(\gamma^i(s_t))$$
$$\gamma^i(s_{t-1}) = \sum_{a_t \in \boldsymbol{A}} \pi^i(a_t \mid s_{t-1}) \exp[-d^i(a_t, s_{t-1})], \tag{1.18}$$

kde

$$d^i(a_t, s_{t-1}) = \sum_{s_t \in \boldsymbol{S}} p(s_t \mid a_t, s_{t-1}) \ln \left[\frac{p(s_t \mid a_t, s_{t-1})}{p^i(s_t \mid a_t, s_{t-1})} \right].$$

Důkaz je k nalezení v [10].

1.4 Bayesovské učení

Návrh optimální politiky předpokládá znalost modelu systému, viz Obrázek 1 a Věta 2 či Věta 3. Prvotní představy o tomto modelu mohou být v průběhu rozhodování zlepšeny učením. Učení vylepšuje model systému za pomocí statistiky V, vícerozměrné funkce naměřených dat, v závislosti na naměřených datech. Tato statistika uchovává výskyty systému ve stavu $\tilde{s}$, za volby akce a a výskytu systému ve stavu s v předešlém kroku. Je zřejmé, že po každém výskytu v $(\tilde{s}, a, s)$ se pravděpodobnost, že se opět systém vyskytne v $(\tilde{s}, a, s)$, zvyšuje.

Jinými slovy, čím vícekrát se systém posune konkrétně ze stavu s do stavu $\tilde{s}$ za volby akce a, tím větší bude pravděpodobnost, že se opět systém ze stavu s posune do stavu $\tilde{s}$ za volby akce a.

Věta 4. *Nechť je dána počáteční hodnota funkce* V, $V_0 > 0$, *a nechť* s_t, a_t, s_{t-1} *jsou konkrétní naměřené hodnoty v čase* $t \in \boldsymbol{T}$. *Poté je v každém čase* $t \in \boldsymbol{T}$ *vypočítána hodnota statistiky* V_t, *takto*

$$V_t(\tilde{s} \mid a, s) = \begin{cases} V_{t-1}(\tilde{s} \mid a, s) & (\tilde{s}, a, s) \neq (s_t, a_t, s_{t-1}) \\ V_{t-1}(\tilde{s} \mid a, s) + 1 & (\tilde{s}, a, s) = (s_t, a_t, s_{t-1}) \end{cases} \tag{1.19}$$

kde $s_t, s_{t-1} \in \boldsymbol{S}, a \in \boldsymbol{A}$.

A model systému je dán vzorcem

$$p(\tilde{s} \mid a, s) \approx p(\tilde{s} \mid a, s, V = V_{t-1}) = \frac{V(\tilde{s} \mid a, s)}{\sum_{\tilde{s} \in \boldsymbol{S}} V(\tilde{s} \mid a, s)}. \tag{1.20}$$

Podrobná konstrukce tohoto tvrzení je k nalezení v [7] a [15].

Kapitola 2

Hledání optimálního ideálu

V této kapitole se hledá optimální ideální distribuce chování uzavřené smyčky, která reflektuje agentovy preference. Jak je již popsáno výše, **PPN** využívá **KL** divergenci distribuce popisující reálné chování k distribuci ideální pro nalezení optimální politiky. Hodnota ideální pravděpodobnosti přechodu do preferovaného stavu by tedy měla být nejvyšší avšak ne rovna 1. V **PPN** je totiž nutné i nežádoucím stavům přiřadit nějakou malou nenulovou hodnotu pravděpodobnosti přechodu. Kdyby totiž byla hodnota pravděpodobnosti přechodu rovna 1, byla by ideální distribuce chování deterministická. Nebylo by pak možné reálnou distribuci chování k té ideální přiblížit a politika nalezena pomocí **PPN** by nebyla optimální. Otázkou ale stále zůstává, jaké hodnoty konkrétně volit pro ideální distribuci. Teorie popsaná v této kapitole je čerpána z [9].

Předpokládá se, že neúplně zadaný popis preferencí vymezuje víceprvkovou množinu

$$\mathbf{C}^i = \left\{c^i(b)\right\}, \tag{2.1}$$

kde c^i je hustota pravděpodobnosti popisující chování $b \in \mathbf{B}$ ideální uzavřené smyčky.
Optimální ideální distribuce chování uzavřené smyčky je pak získaná minimalizací minim (1.14) přes $\mathbf{C}^i$:

$$c_0^i \in \operatorname{Arg}\min_{c^i \in \mathbf{C}^i} \min_{\pi \in \mathbf{\Pi}} D(c^\pi \| c^i). \tag{2.2}$$

Pomocí řetězového pravidla lze opět c_0^i zapsat jako

$$c_0^i(b) = \prod_{t \in \mathbf{T}} p_0^i(s_t \mid a_t, s_{t-1}) \pi_0^i(a_t \mid s_{t-1}). \tag{2.3}$$

Proto může být a je optimalizace c_0^i provedena pomocí optimálního ideálního modelu systému p_0^i a optimálního ideálního rozhodovacího pravidla π_0^i. Rovnice (1.17), (1.18), (2.2) implikují

$$c_0^i \in \operatorname{Arg}\min_{c^i \in \mathbf{C}^i} \min_{\pi \in \mathbf{\Pi}} D(c^\pi \| c^i) = \operatorname{Arg}\min_{c^i \in \mathbf{C}^i} [-\ln(\gamma(s_0))] \leq \operatorname{Arg}\min_{c^i \in \mathbf{C}^i} [-\ln(\gamma^i(s_0))]. \tag{2.4}$$

Minimalizace $[-\ln(\gamma(s_0))]$ přes c^i je většinou nedosažitelná, zato $\operatorname{Arg}\min_{c^i \in \mathbf{C}^i} [-\ln(\gamma^i(s_0))]$ lze vypočítat analyticky. Aproximace modelu optimální ideální uzavřené smyčky je tedy nalezena pomocí minimalizace horní hranice. Jelikož je ale tato minimalizace přes $c^i = p^i \pi^i$ formálně identická pro všechny časy $t \in \mathbf{T}$ a všechny stavy s_{t-1}, lze potlačit označení času t a stavu s_{t-1} a tuto minimalizaci vyjádřit obecně jako

$$c_0^i \in Arg\min_{c^i \in \mathbf{C}^i} [-\ln(\gamma^i)]. \tag{2.5}$$

Tato minimalizace je ekvivalentní maximalizaci γ^i v (1.18) přes (2.1). Proto pro $c^i(s,a) = p^i(s \mid a)\pi^i(a)$ a $d^i(a)$ definovaném v (1.18) je

$$c_0^i = p_0^i \pi_0^i \in \operatorname{Arg}\max_{p^i \in \mathbf{P}^i} \left[\max_{\pi^i \in \mathbf{\Pi}^i} \sum_{a \in \mathbf{A}} \pi^i(a) \exp[-d^i(a)] \right]. \tag{2.6}$$

Optimální c_0^i je pak nalezeno přes dvě maximalizace (přes π^i a p^i).

2.1 Optimalizace přes π^i

Pro optimalizaci přes π^i je nutno specifikovat množinu $\mathbf{\Pi}^i$ ideálních rozhodovacích pravidel tak, aby dodržovaly agentovy preference.

Nosič přípustných rozhodovacích pravidel je $\operatorname{supp}[\pi] = \{a : \pi(a) > 0\} \subseteq \mathbf{A}$. Tvar optimálního rozhodovacího pravidla π^0 (1.15) implikuje $\operatorname{supp}[\pi] \subseteq [\pi^i]$. **PPN** přiřazuje nenulovou pravděpodobnost všem možným akcím v $\operatorname{supp}[\pi^i]$. Proto musí ideální rozhodovací pravidlo splňovat

$$\pi^i \in \mathbf{\Pi}^i = \{\pi^i : \operatorname{supp}[\pi^i] = \mathbf{A}\}. \tag{2.7}$$

Pro dané p^i lze rovnici maximalizace (2.6) přepsat takto

$$\begin{aligned} \max_{\pi^i \in \mathbf{\Pi}^i} \gamma^i &= \max_{\pi^i \in \mathbf{\Pi}^i} \sum_{a \in \mathbf{A}} \pi^i(a) \exp(-d^i(a)) \\ d^i(a) &= \sum_{s \in \mathbf{S}} p(s \mid a) \ln\Big(\frac{p(s \mid a)}{p^i(s \mid a)}\Big). \end{aligned} \tag{2.8}$$

Nutnou podmínkou smysluplnosti maximalizace (2.8) je zřejmě konečnost $d^i(a)$. Optimalizované π^i vstupuje do γ^i lineárně. Kdyby bylo γ^i maximalizováno bez omezení, pak by se rozhodovací pravidlo π_0^i soustřeďovalo kolem hodnoty akce rovné $\arg\min_{a \in \mathbf{A}} d^i(a)$, ale takové rozhodovací pravidlo by nesplňovalo (2.7). Proto musí být zavedeno omezení.

Věta 5. *Nechť $\boldsymbol{A}$ je podmnožina konečně dimenzionálního reálného prostoru a π je hustota pravděpodobnosti na $\boldsymbol{A}$. Definujeme*

$$k = \begin{cases} 1 & |\boldsymbol{A}| < \infty \\ \infty & |\boldsymbol{A}| = \infty \end{cases}$$

Pak platí

$$k = \sum_{a \in \boldsymbol{A}} \pi^2(a) \iff \pi(a) \text{ je deterministické.} \tag{2.9}$$

Tudíž dodatečná podmínka na π^i zabraňující rovnosti (2.9) umožní, aby π_0^i maximalizující γ^i bylo z (2.7). Následující věta využívá indikátoru

$$\chi_{\mathbf{A}}(a) = \begin{cases} 1 & a \in \mathbf{A} \\ 0 & a \notin \mathbf{A} \end{cases} \tag{2.10}$$

Věta 6. *Nechť*

$$d^i(a) = \sum_{s \in \boldsymbol{S}} p(s \mid a) \ln\Big(\frac{p(s \mid a)}{p^i(s \mid a)}\Big) < \infty. \tag{2.11}$$

To je ekvivalentní implikaci

$$p^i(s \mid a) = 0 \Rightarrow p(s \mid a) = 0 \text{ na } \boldsymbol{S}, \tag{2.12}$$

*což vyplývá z vlastnosti **KL** divergence.*

Pak π_0^i maximalizující γ^i z (2.8) přes $\boldsymbol{\Pi}^i$ (2.7), parametrizované $\varkappa \in (0, k)$,

$$\boldsymbol{\Pi}^i = \left\{\pi^i(a) : \sum_{a \in \boldsymbol{A}} (\pi^i)^2(a) \leq \varkappa < k\right\} \tag{2.13}$$

má tvar

$$\pi_0^i \propto \chi_{\boldsymbol{A}}(a) \exp[-d^i(a)]. \tag{2.14}$$

Takto optimalizované ideální rozhodovací pravidlo má $supp[\pi^i] = \boldsymbol{A}$, a tudíž splňuje (2.7).

Důkazy těchto tvrzení jsou k nalezení v [9], odkud jsem čerpala i danou teorii.

2.2 Optimalizace přes p^i

Požadavek (2.13) na π^i je dost obecný, ostatní preference mohou být více závislé na jednotlivých případech.

Citovaná práce [9] vyjadřuje agentovy preference l_q-vektorovou funkcí $q(s, a), s \in \mathbf{S}, a \in \mathbf{A}$. Tato funkce definuje omezení na momenty chování v obvodu tvořeném daným modelem systému a ideálním rozhodovacím pravidlem

$$0 = \sum_{a \in \mathbf{A}} \sum_{s \in \mathbf{S}} q(s, a) p(s \mid a) \pi^i(a), \tag{2.15}$$

kde $p(s \mid a)$ je pevný model systému a π^i ideální rozhodovací pravidlo. Pevný stav s_{t-1} v podmínce je zde stále potlačen. Po dosazení optimálního rozhodovacího pravidla π_0^i (2.14) do (2.15) vzniká omezení na ideální model systému p^i

$$0 = \sum_{a \in \mathbf{A}} \sum_{s \in \mathbf{S}} q(s, a) p(s \mid a) \times \exp\left[-\sum_{s \in \mathbf{S}} p(\tilde{s} \mid a) \ln\left(\frac{p(\tilde{s} \mid a)}{p^i(\tilde{s} \mid a)}\right)\right]. \tag{2.16}$$

Typicky je agentova preference vyjádřena funkcí $q(s, a) = s - s^i$, tj. je požadováno, aby se stav $s \in \mathbf{S}$ stabilizoval v preferovaném stavu $s^i \in \mathbf{S}$.

Věta 7. *Nechť je sada ideálních uzavřených smyček $\boldsymbol{C}^i = (\boldsymbol{\Pi}^i, \boldsymbol{P}^i)$ dána (2.13), (2.11), (2.15) neprázdná.*[1] *Pak $p_0^i(s \mid a)$ maximalizující γ^i (2.8) přes $(p^i, \pi_0^i) \in \boldsymbol{C}^i$ je ($'$ je transpozice)*

$$p^i(s \mid a) = \frac{p(s \mid a) \exp(\lambda' q(s, a))}{\sum_{s \in \boldsymbol{S}} p(s \mid a) \exp(\lambda' q(s, a))}, \tag{2.17}$$

kde l_q-vektor λ je takový, aby platilo (2.16).

Důkaz tvrzení je opět k nalezení v [9].

[1] Existence p^i splňující (2.12), (2.16) garantuje neprázdnost

2.3 Optimální c_0^i

Optimální ideální uzavřená smyčka c_0^i maximalizující γ^i (2.8) a tím pádem minimalizující horní hranici hodnoty funkce (1.18) přes sadu ideálních uzavřených smyček $\mathbf{C}^i$ definovanou (2.11), (2.13), (2.15) je

$$c_0^i(s,a) = p_0^i(s \mid a)\pi_0^i(a) \propto \frac{p(s \mid a)\exp[\lambda' q(s,a)]}{\sum_{s\in\mathbf{S}} p(s \mid a)\exp[\lambda' q(s,a)]} \times \chi_{\mathbf{A}}(a)\frac{\exp[\sum_{s\in\mathbf{S}} \lambda' q(s,a)p(s \mid a)]}{\sum_{s\in\mathbf{S}} p(s \mid a)\exp[\lambda' q(s,a)]}. \tag{2.18}$$

l_q-vektor λ v (2.18) je řešením rovnice

$$0 = \sum_{a\in\mathbf{A}}\sum_{s\in\mathbf{S}} q(s,a)p(s \mid a)\pi_0^i(a).$$

Řešitelnost této rovnice se předpokládá. Tento předpoklad je splnitelný pro spojité stavy a akce, avšak často porušen v případě uvažovaných diskrétních stavů a akcí. Tento problém je řešen v (sekce 2.5). Navržené obecnější řešení je jedním z přínosů této práce.

2.4 Přidání preference na akce

Optimalizace ideálního rozhodovacího pravidla, která je uvedená výše, sice maximalizuje γ^i, ale nezahrnuje žádnou preferenci na hodnoty akce. Proto je žádoucí přidat ještě dodatečnou podmínku na volbu $a \in \mathbf{A}$. Jednou z těchto preferencí je například volba malých $a \in \mathbf{A}$ nebo naopak volba velkých a. Agent může preferovat menší akce především kvůli vynaložení menší energie volbou menších akcí k posunu do dalšího stavu. Tuto preferenci vysvětlím na triviálním příkladu pro lepší pochopení.

Příklad V zimě agent preferuje v místnosti vysokou teplotu (preference stavu), ale nechce příliš topit (preference akce).
Kdyby měla být splněna preference "nechci příliš topit" bez omezení, pak by se také mohlo stát, že by se netopilo vůbec, tím by ale nebyla splněna preference stavu "chci vysokou teplotu". Proto je zapotřebí preferenci ohodnotit váhou, která bude vyjadřovat, jak moc tuto preferenci agent vyžaduje s ohledem na preferenci stavu.

Intuitivně by agentova preference na akce měla být vyjádřena nějakou funkcí akcí, která by vyjadřovala danou preferenci, a váhou, což by byl kladný parametr, který by porovnával, jak moc agent vyžaduje danou preferenci s ohledem na preferenci stavu. Parametr váhy se označí ω a bude splňovat $0 \leq \omega \leq 1$. Funkce vyjadřující preferenci se označí $\phi(a)$. Funkce $\phi(a)$ v této práci vyjadřuje preferenci co nejmenších akcí.

Pro maximalizaci γ^i, a také z důvodu řešení tohoto problému na bázi pravděpodobnosti, je ale nutné udržet celou ideální hustotu pravděpodobnosti kladnou, proto se zdá jako nejlepší vyjádřit ji pomocí exponenciály. Ideální rozhodovací pravidlo s přidanou preferencí na akce je pak tvaru

$$\pi^i(a) \propto \tilde{\pi}^i(a)\exp(-\omega\phi(a)) \tag{2.19}$$

a maximalizace γ^i vypadá takto

$$\max_{\tilde{\pi}^i\in\mathbf{\Pi}^i} \gamma^i = \max_{\tilde{\pi}^i\in\mathbf{\Pi}^i} \sum_{a\in\mathbf{A}} \tilde{\pi}^i(a)\exp(-d^i(a) - \omega\phi(a)). \tag{2.20}$$

Pro preferenci co nejmenších akcí funkce $\phi(a)$ přiřazuje malé kladné hodnoty požadovaným a velké kladné hodnoty nežádoucím akcím, jelikož budou nejčastěji vybírány akce s malou funkční hodnotou $\phi(a)$ z důvodu maximalizace $\exp(-\omega\phi(a))$. Na první pohled je vidět, že se jedná o kladnou funkci, která je velmi podobná (2.8). Proto bude i volba podmínek velmi podobná předchozímu řešení optimalizace π^i.

Hodnota k bude totožná jako v (2.9) a pro splnění (2.7) se opět využije indikátor $\chi_{\mathbf{A}}(a)$ z (2.10). Posledním požadavkem pro nalezení optima je konečnost argumentu exponenciály. Jak už je uvedeno výše (2.11), $d^i(a)$ je omezenou funkcí ($0 \leq d^i(a) < \infty$). Zbývá tedy určit podmínky pro volbu $\omega\phi(a)$.

ω je zde pouze kladná konstanta mezi 0 a 1, takže podmínky směřují pouze na $\phi(a)$. $\phi(a)$ musí být zvolena konečná $\phi(a) < \infty$ a zároveň musí být splněno (2.7). Proto je nutné, aby byla $\phi(a)$ definovaná pro všechna $a \in \mathbf{A}$.

Věta 8. *Nechť jsou splněny stejné podmínky (2.8), (2.9), (2.10), (2.11), (2.13) jako pro optimalizaci π_0^i (2.14) a dále at' $0 \leq \omega \leq 1$ je parametr váhy, (pro $\omega = 0$ řešení splývá s (2.14)) a nechť funkce akcí splňuje $0 \leq \phi(a) < \infty$ a je definovaná $\forall a \in \mathbf{A}$.*

Pak je optimální rozhodovací pravidlo tvaru

$$\pi_0^i \propto \chi_{a\in\mathbf{A}}(a) \exp(-d^i(a) - \omega\phi(a)). \tag{2.21}$$

Důkaz. Jako v předchozí optimalizaci ekvivalence konečnosti $d^i(a)$ (2.11) a implikace (2.12) vychází z vlastnosti **KL** divergence. Díky indikátoru $\chi_{a\in\mathbf{A}}(a)$ a (2.12) je ideální rozhodovací pravidlo z (2.7). Bez omezení dosáhne $\tilde{\pi}^i \exp(-\omega\phi(a))$ maximalizující γ^i hranici k (2.9). Za pomocí omezení (2.13) je dosaženo pouze hranice

$$\sum_{a\in\mathbf{A}} \left(\tilde{\pi}^i \exp(-\omega\phi(a))\right)^2 (a) = \varkappa \in (0, 1). \tag{2.22}$$

Poté díky konečnosti $d^i(a)$, kladnosti $\exp[-d^i(a)]$ a omezujících podmínek na $\phi(a)$, je

$$\beta = \sum_{a\in\mathbf{A}} \left(\chi_{a\in\mathbf{A}}(a) \exp[-d^i(a)]\right)^2 \tag{2.23}$$

kladná a konečná. Pak normované γ^i

$$\frac{\sum_{a\in\mathbf{A}} \tilde{\pi}^i(a)\chi_{a\in\mathbf{A}}(a) \exp[-d^i(a) - \omega\phi(a)]}{\sqrt{\varkappa\beta}} \tag{2.24}$$

může být maximalizováno místo původního γ^i. Maximalizované (2.24) je skalárním součinem funkcí $\tilde{\pi}^i \exp[-\omega\phi(a)]$ a $\chi_{a\in\mathbf{A}} \exp[-d^i(a))]$ děleným součinem jejich norem.

$$\max_{||\tilde{\pi}^i \exp[-\omega\phi(a)]|| \leq \sqrt{\varkappa}} \left\langle \frac{\tilde{\pi}^i \exp[-\omega\phi(a)]}{\sqrt{\varkappa}}, \frac{\chi_{a\in\mathbf{A}} \exp[-d^i(a)]}{||\chi_{a\in\mathbf{A}} \exp[-d^i(a)]||} \right\rangle. \tag{2.25}$$

Hodnota (2.24) je kosinus úhlu mezi těmito dvěma funkcemi. Z čehož vyplývá, že tato hodnota bude maximální, pokud funkce budou kolineární. Což vyjadřuje (2.21) □

2.5 Obecnější optimalizace p^i

Optimalizace p^i za podmínky na momenty (2.15), která byla sestrojena původně pro spojité stavy a akce, se bohužel pro diskrétní stavy a akce nehodí. V diskrétním světě může být množina řešení často

prázdná. Neprázdnost množiny řešení tedy není v diskrétním světě garantována a je potřeba nalézt obecnější formulaci optimalizace p^i.

Obecná optimalizace ideálního modelu systému vychází z podmínky maximalizace γ^i (2.8). Dále se požaduje, aby optimální ideální model systému vystihoval agentovy preference. Jelikož zejména model systému určuje výskyt systému v dalším stavu, je nutné, aby pravděpodobnost dosažení preferovaného stavu $s^i \in \mathbf{S}$ byla větší nebo rovna pravděpodobnosti dosažení stavu jiného. To vyjadřuje podmínka

$$p^i(s^i \mid a) \geq p^i(s \mid a) \tag{2.26}$$

pro všechna $s \in \mathbf{S}$ a pro $a \in \mathbf{A}$ pevné.

Věta 9. *Nechť je dána sada uzavřených smyček $\boldsymbol{C}^i = (\boldsymbol{\Pi}^i, \boldsymbol{P}^i)$ a nechť s^i označuje preferovaný stav. Pak $p_0^i(s \mid a)$ maximalizující γ^i (2.8) přes $(p^i, \pi_0^i) \in \boldsymbol{C}^i$ za podmínky $p(s^i \mid a) \geq p(s \mid a)$, je rovno*

$$p^i(s \mid a) = p(s \mid a), \tag{2.27}$$

pokud (2.27) splňuje (2.26) pro pevně dané $a \in \boldsymbol{A}$, $\forall s \in \boldsymbol{S}$.

Jinak pokud existuje $\tilde{s}$ takové, že $p(\tilde{s} \mid a) > p(s^i \mid a)$, pak

$$p^i(s^i \mid a) = \max_{s \in \mathbf{S}} \frac{p(s \mid a)}{1 + p(s \mid a) - p(s^i \mid a)} \tag{2.28}$$

a pro $s \neq s^i$

$$p^i(s \mid a) = \frac{1 - p^i(s^i \mid a)}{1 - p(s^i \mid a)} p(s \mid a). \tag{2.29}$$

Důkaz. Tato optimalizace vychází z požadavku maximalizace

$$\sum_{a \in \mathbf{A}} \pi^i(a) \exp[-d^i(a)],$$

kde se po dosazení optimálního ideálního rozhodovacího pravidla

$$\pi_0^i = \chi_{\mathbf{A}}(a) \exp[-d^i(a)],$$

získá

$$\max_{p^i} \sum_{a \in \mathbf{A}} \chi_{\mathbf{A}}(a) \exp[-2d^i(a)].$$

Aby byla tato funkce maximální, je potřeba volit $d^i(a)$ (2.8) minimální

$$d^i(a) = \sum_{s \in \mathbf{S}} p(s \mid a) \ln\left(\frac{p(s \mid a)}{p^i(s \mid a)}\right),$$

a to za podmínky (2.27)

$$p^i(s^i \mid a) \geq p^i(s \mid a),$$

$\forall s \in \mathbf{S}$.

Z této podmínky hned jednoduše vyplývá, že pokud pro model systému platí, že $p(s^i \mid a) \geq p(s \mid a)$ pro pevně dané $a \in \mathbf{A}$ a $\forall s \in \mathbf{S}$, pak stačí ideální model systému položit roven

$$p^i(s \mid a) = p(s \mid a).$$

Pokud ale existuje $\tilde{s}$ takové, že $p(\tilde{s} \mid a) > p(s^i \mid a)$ je nutno pokračovat dále v minimalizaci. Pro zjednodušení se označí $p^i(s^i \mid a) = \alpha$.

Jelikož minimalizace $d^i(a)$ v α nezávisí na čitateli v argumentu logaritmu, je čitatel argumentu logaritmu vzhledem k α konstantní, lze označit $\sum_{s \in \mathbf{S}} p(s \mid a) \ln(p(s \mid a)) = \tilde{C}$.

Dále pomocí jednoduchých matematických operací

$$\begin{aligned} \sum_{s \in \mathbf{S}} p(s \mid a) \ln\Big(\frac{1}{p^i(s \mid a)}\Big) + \tilde{C} = p(s^i \mid a) \ln\Big(\frac{1}{\alpha}\Big) + \sum_{s \neq s^i} p(s \mid a) \ln\Big(\frac{1}{\frac{p^i(s|a)(1-\alpha)}{1-\alpha}}\Big) + \tilde{C} = \\ -p(s^i \mid a) \ln \alpha - \Big(1 - p(s^i \mid a)\Big) \ln(1-\alpha) + \Big(1 - p(s^i \mid a)\Big) \sum_{s \neq s^i} \frac{p(s \mid a)}{1 - p(s^i \mid a)} \ln\Big(\frac{1}{\frac{p^i(s|a)}{1-\alpha}}\Big) + \tilde{C}. \end{aligned} \tag{2.30}$$

Suma

$$\sum_{s \neq s^i} \frac{p(s \mid a)}{1 - p(s^i \mid a)} \ln\Big(\frac{1}{\frac{p^i(s|a)}{1-\alpha}}\Big)$$

je minimální, pokud

$$\frac{p(s \mid a)}{1 - p(s^i \mid a)} = \frac{p^i(s \mid a)}{1 - \alpha}, \tag{2.31}$$

$\forall s \in \mathbf{S} \setminus \{s^i\}$, jak plyne z (Věta 1).

Po vyjádření $p^i(s \mid a)$ z (2.31) a po dosazení do (2.27) se získá omezení pro α

$$1 \geq \alpha \geq \frac{(1-\alpha)p(s \mid a)}{1 - p(s^i \mid a)} \geq 0, \tag{2.32}$$

$\forall s \in \mathbf{S} \setminus \{s^i\}$. Hodnoty $\alpha = 1$ a $\alpha = 0$ nejsou optimální, jelikož ideální model systému nesmí být deterministický, posun do každého stavu musí mít nenulovou pravděpodobnost.

Podmínky na α jsou $\alpha \in (0, 1)$ a současně z (2.32) plyne

$$\alpha \geq \frac{p(s|a)}{1 + p(s|a) - p(s^i|a)}. \tag{2.33}$$

Nejdříve se provede minimalizace (2.30) bez omezení.

$$\frac{\partial}{\partial \alpha} : -p(s^i \mid a)\frac{1}{\alpha} - \Big(1 - p(s^i \mid a)\Big)\frac{-1}{1-\alpha} = 0$$

Za podmínek $\alpha \neq 1$ a $\alpha \neq 0$ lze rovnici vynásobit $-\alpha(1-\alpha)$.

$$p(s^i \mid a)(1-\alpha) - \Big(1 - p(s^i \mid a)\Big)\alpha = 0,$$

pak $\alpha = p(s^i \mid a)$.

Jelikož je druhá derivace nezáporná

$$\frac{\partial^2}{\partial \alpha^2} : p(s^i \mid a)\frac{1}{\alpha^2} + \Big(1 - p(s^i \mid a)\Big)\frac{1}{(1-\alpha)^2} \geq 0,$$

nastává v bodě α minimum. Takové α sice leží v intervalu $(0, 1)$, ale nesplňuje $\alpha \geq \frac{p(s|a)}{1+p(s|a)-p(s^i|a)}$. Omezení tedy musí být aktivní a v podmínce (2.33) nastane rovnost.

Z toho vyplývá, že pokud pro pevné $a \in \mathbf{A}$ existuje $\tilde{s}$ takové, že $p(\tilde{s} \mid a) > p(s^i \mid a)$ minimum $d^i(a)$ nastává v bodě

$$\alpha = p^i(s^i \mid a) = \max_{s \in \mathbf{S}} \frac{p(s \mid a)}{1 + p(s \mid a) - p(s^i \mid a)}$$

a pro $s^i \neq s$, $s \in \mathbf{S}$ se ideál vypočítá z (2.31) jako

$$p^i(s \mid a) = \frac{1 - p^i(s^i \mid a)}{1 - p(s^i \mid a)} p(s \mid a).$$

Tímto byly obě hodnoty α pro minimum vypočítány a věta tak byla dokázána. □

Tato optimalizace má vždy řešení a je tak obecná, že lze analogicky využít jak v diskrétním, tak spojitém světě.

Kapitola 3

Experimenty

V této kapitole je několik experimentálních úloh, které ilustrují teorii popsanou výše. Experimenty jsou rozděleny do dvou skupin. První se zabývá vlivem učení modelu systému a druhá optimalizovaným ideálem. Tyto úlohy mají za úkol porovnat pevně dané hodnoty zvnějšku s optimalizovanými hodnotami. Všechny tyto úlohy jsou zpracované v Matlabu.

3.1 Simulace systému

V této sekci jsou zvoleny parametry, které jsou pro většinu experimentů pevné.

$|\mathbf{S}|$, $|\mathbf{A}|$ je počet stavů/akcí, s_0 je počáteční stav, s^i je preferovaný stav, h je horizont, $d_{simulace}$ je délka simulace, ω je váha pro přidání preference na akce, pro počáteční experimenty však nebude kladená žádná preference na akce, proto $\omega = 0$. Seed je funkce Matlabu, která nastaví počáteční podmínky experimentů tak, aby všechny experimenty měly stejné počáteční podmínky. Pokud se simulace pustí s určitým seedem několikrát, jsou vždy generovány stejné pseudonáhodné posloupnosti. Tato funkce je použita bez újmy na obecnosti a je využita proto, aby výsledky experimentů byly porovnatelné.

$\|\mathbf{S}\|$=3	$\|\mathbf{A}\|$=3
$s_0 = 1$	s^i=1
$h = 100$	$d_{simulace}$=500
$\omega = 0$	seed=0

Tabulka 3.1: Pevně zvolené parametry

A $\phi(a)$ je funkce vyjadřující preference akcí (2.21). Programátorsky byla $\phi(a)$ označena "phi".

```
phi(1)=1 phi(2)=2 phi(3)=3
```

Což odpovídá preferenci malých hodnot akcí.

Pravděpodobnosti přechodu systému zvolené pro základní sadu experimentů byly označeny P_0. Simulovaný systém byl zvolen tak, aby výrazně ukazoval sledované vlastnosti. Na něm jsou výsledky vidět nejvýrazněji a dost čitelně, proto byl pro lepší pochopení teorie vybrán právě tento systém. Experimenty byly samozřejmě vyzkoušeny na řadě systémů, u některých byly výsledky stejně výrazné jako u tohoto systému a u jiných byly výsledky horší. Je jasné, že pro nějaké systémy optimalizace nikdy nebude dost viditelná a nebude vyhovovat preferencím agenta, jelikož systém upřednostňuje jiné stavy. Je tomu tak i v reálném světě, proto nelze ve všech systémech dosáhnout agentových preferencí, jen se k nim přiblížit. Programátorsky bylo P_0 označeno "P_0".

```
P_0=zeros(|S|,|A|,|S|)
P_0(1,1,1)=0.90  P_0(2,1,1)=0.05  P_0(3,1,1)=0.05
P_0(1,2,1)=0.05  P_0(2,2,1)=0.90  P_0(3,2,1)=0.05
P_0(1,3,1)=0.05  P_0(2,3,1)=0.05  P_0(3,3,1)=0.90

P_0(:,:,3)=P_0(:,:,2)=P_0(:,:,1)
```

Pokud nebylo použito bayesovské učení, pravděpodobnosti přechodu modelu systému byly zvoleny rovnoměrně. Programátorsky byl model systému označen "p".

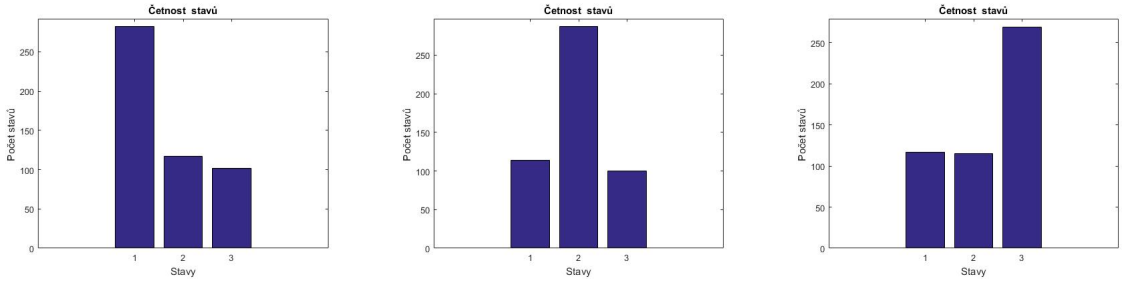
```
p=ones(|S|,|A|,|S|)/|S|
p(1,1,1)=0.3333  p(2,1,1)=0.3333  p(3,1,1)=0.3333
p(1,2,1)=0.3333  p(2,2,1)=0.3333  p(3,2,1)=0.3333
p(1,3,1)=0.3333  p(2,3,1)=0.3333  p(3,3,1)=0.3333

p(:,:,3)=p(:,:,2)=p(:,:,1)
```

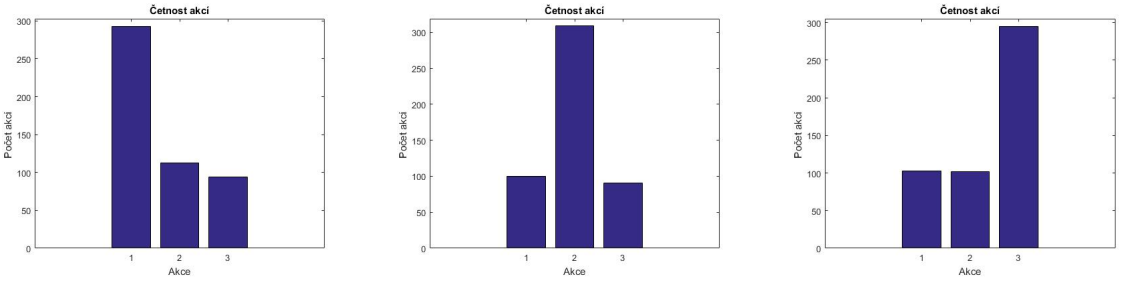
3.2 Vliv bayesovského učení

Experiment 1. První úloha slouží jako srovnávací. Ukazuje, jaké kvality rozhodování by bylo možno dosáhnout s daným systémem při jeho úplné znalosti.

Obrázek 3.1: Experiment 1.



(1) Četnosti jednotlivých stavů s úplnou znalostí systému, $s^i = 1$ (2) Četnosti jednotlivých stavů s úplnou znalostí systému, $s^i = 2$ (3) Četnosti jednotlivých stavů s úplnou znalostí systému, $s^i = 3$



(4) Četnosti jednotlivých akcí s úplnou znalostí systému, $s^i = 1$ (5) Četnosti jednotlivých akcí s úplnou znalostí systému, $s^i = 2$ (6) Četnosti jednotlivých akcí s úplnou znalostí systému, $s^i = 3$

Statistika V, z výše popsaného učení (sekce 1.4), byla zvolena rovna $1000P_0$, kde P_0 je matice pravděpodobnosti přechodu systému. Tím, že se hodnoty P_0 vynásobily 1000, bylo učení modelu zanedbatelné (k V se v každém kroku přičítala pouze 1, takže data neměla skoro žádný vliv). Tato úloha byla pro

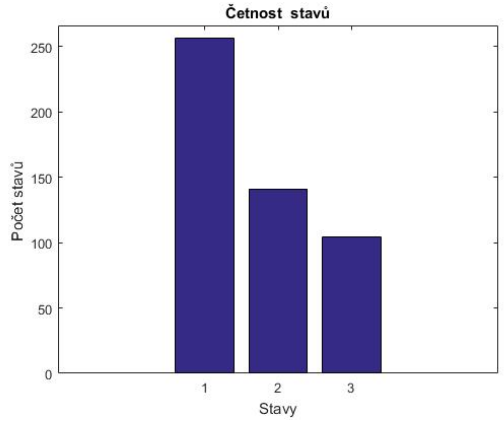
jasnější představu vykreslena pro všechny možné preference stavů.

Diskuze Jak je patrné z grafů (Experiment 1.), preference jsou splněny pro všechny tři preferované stavy. Preferovaný stav je v simulaci o délce 500 časových epoch zastoupen více než dvojnásobně v porovnání s dalšími dvěma stavy. Jak je již zmíněno výše, úplná znalost systému umožňuje nejlépe optimalizovat agentova rozhodovací pravidla, protože agent přesně zná model systému. To není v reálném světě samozřejmě realizovatelné.

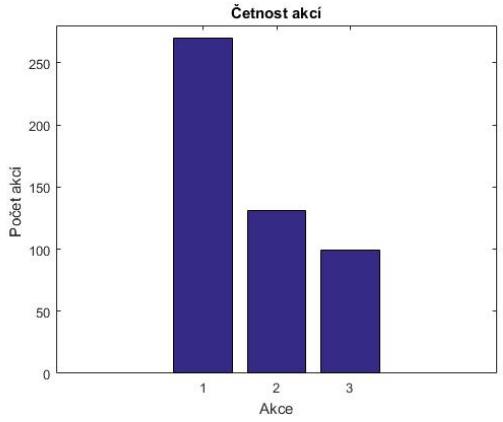
Další úlohy jsou zaměřeny na více reálné předpoklady. Je ale dobré si povšimnout, že ač simulovaný systém nemá paměť (pravděpodobnost stavu s_t závisí na a_t a nikoliv na s_{t-1}) všechny tři hodnoty akcí jsou užity, jak odpovídá požadavku $supp[\pi^i(a \mid s)] = \mathbf{A}$.

Experiment 2. Druhý experiment navazuje na experiment předchozí. V tomto experimentu však nebyl znám úplný model systému, ale postupně se počítal pomocí bayesovského učení (sekce 1.4).

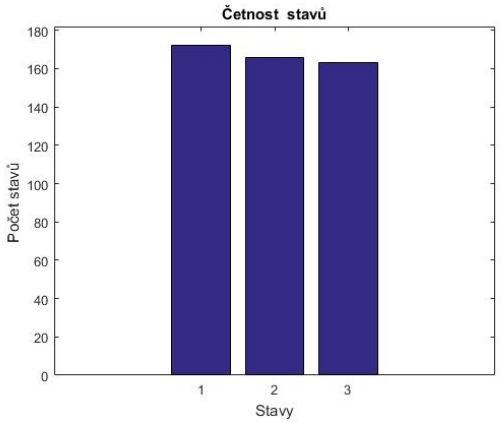
Obrázek 3.2: Experiment 2.



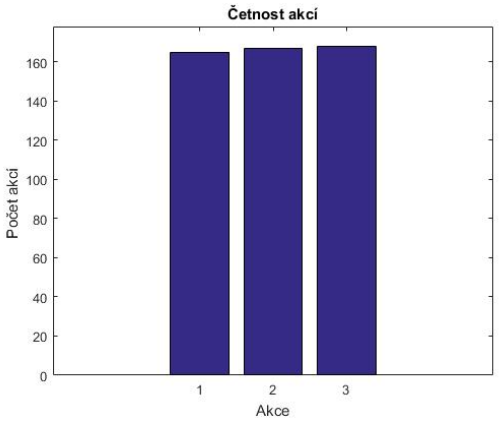
(1) Četnosti jednotlivých stavů s optimalizací a s učením, $s^i = 1$



(2) Četnosti jednotlivých akcí s optimalizací a s učením, $s^i = 1$



(3) Četnosti jednotlivých stavů s optimalizací, bez učení, $s^i = 1$



(4) Četnosti jednotlivých akcí s optimalizací, bez učení, $s^i = 1$

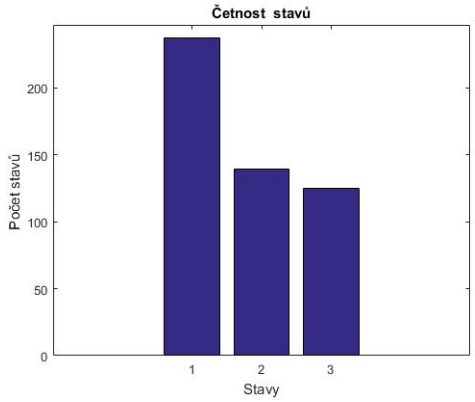
Byla zde porovnána simulace s optimalizací a s učením se simulací s optimalizací bez učení. Počáteční hodnota všech prvků V byla rovna 1, což odpovídá nerealistickému (ale málo fixovanému) modelu

systému, který očekává stejnou pravděpodobnost všech přechodů nezávislou na zvolené akci. Postupně se V opravovala pozorovanými daty, tj. model systému se opravoval pomocí učení. V tomto experimentu by mělo být očividné, že při učení dochází ke splnění preference výrazně častěji, jelikož agent už má nějakou znalost přechodových pravděpodobností systému a může tak vyvodit, při jaké akci se posune do preferovaného stavu s větší pravděpodobností.

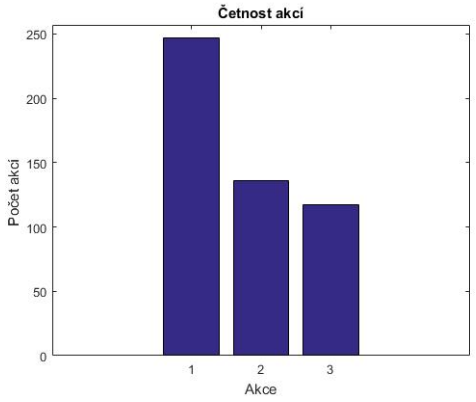
Diskuze V této úloze je model bez učení volen rovnoměrně, tedy všechny stavy mají stejnou pravděpodobnost, proto samotná optimalizace nestačí (Experiment 2.). Sice je preferovaný stav zastoupen nejčetněji, ale jde o náhodný malý rozdíl. Na grafu akcí bez učení je ještě lépe vidět, že akce jsou voleny rovnoměrně. Na grafech s učením je na druhou stranu na první pohled vidět, že preferovaný stav je zastoupen nejčetněji. Učení tedy optimalizaci podporuje a dává tak mnohem výraznější výsledky. V porovnání s úlohou s úplnou znalostí je četnost preferovaného stavu nižší, což se předpokládalo, jelikož na začátku experimentu nemá agent žádnou znalost systému a počítá si ji sám postupně z předchozích výsledků. Je ale stále viditelné, že optimalizaci s učením opravdu funguje.

Experiment 3. Třetí experiment byl zvolen stejně jako předchozí pouze s jinými počátečními podmínkami.

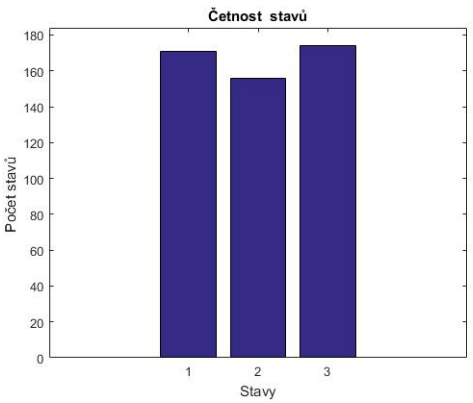
Obrázek 3.3: Experiment 3.



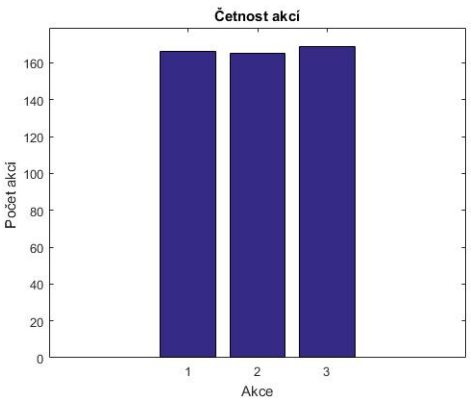
(1) Četnosti jednotlivých stavů s optimalizací a s učením, $s^i = 1$



(2) Četnosti jednotlivých stavů s optimalizací a s učením, $s^i = 1$



(3) Četnosti jednotlivých stavů s optimalizací bez učení, $s^i = 1$



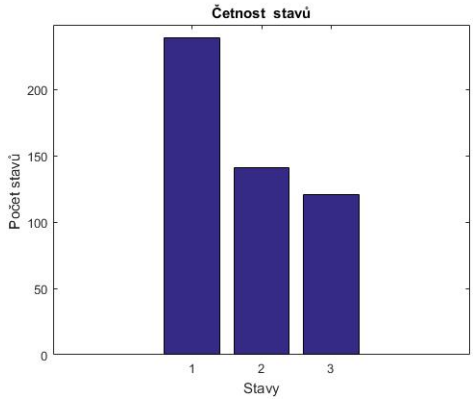
(4) Četnosti jednotlivých stavů s optimalizací bez učení, $s^i = 1$

Tento experiment slouží jako porovnání, jestli jinak náhodně zvolené počáteční podmínky, nezmění vlastnosti experimentu. Pro tento experiment byl pouze změněn *seed* v programu Matlab z 0 na 1, tím se počáteční podmínky upevnily na jinou hodnotu.

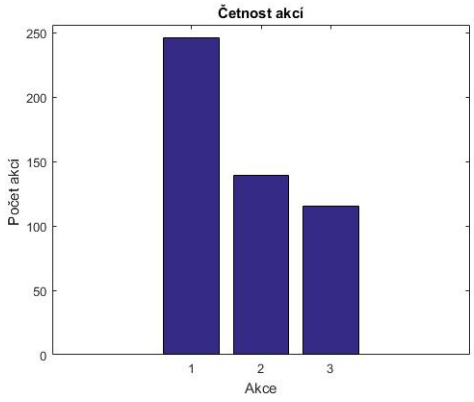
Diskuze Jak je patro z grafů (Experiment 3.), charakter experimentu se nezměnil. Naznačuje to, že tyto experimenty budou dávat výsledky stejného typu pro všechny možné počáteční podmínky. Experimenty jsou tedy prováděné bez újmy na obecnosti. Optimalizace s učením opět dosahuje mnohem lepších výsledků. Rozdíly mezi četnostmi jednotlivých stavů jsou daleko výraznější než s optimalizací bez učení. Učení tedy opět zvýrazňuje charakter optimalizace.

Experiment 4. Čtvrtý experiment by měl být spíše ilustrativní. Porovnávaly se četnosti stavů a akcí bez optimálního rozhodovacího pravidla, tedy s akcemi generovanými rovnoměrně, a s optimálním rozhodovacím pravidlem s využitím učení. Optimální rozhodovací pravidlo volilo akce na základě **PPN** a optimalizovaného ideálu popsaného výše v teorii (sekce 2.4 a 2.5), stejně jako v předchozích experimentech.

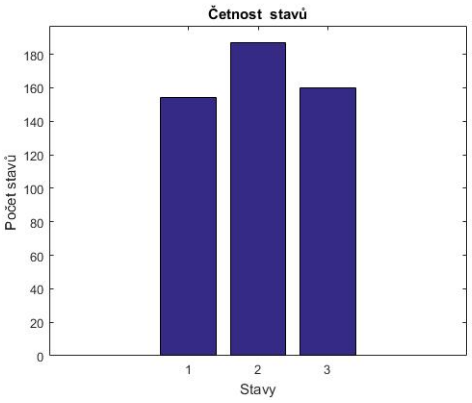
Obrázek 3.4: Experiment 4.



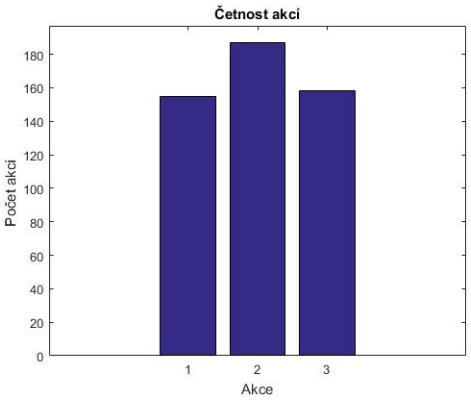
(1) Četnosti jednotlivých stavů s optimalizovaným rozhodovacím pravidlem, $s^i = 1$



(2) Četnosti jednotlivých akcí s optimalizovaným rozhodovacím pravidlem, $s^i = 1$



(3) Četnosti jednotlivých stavů s akcemi generovanými rovnoměrně, $s^i = 1$



(4) Četnosti jednotlivých akcí s akcemi generovanými rovnoměrně, $s^i = 1$

Diskuze Na první pohled je patrno, že pokud agent volí akce náhodně rovnoměrně, nedosáhne preferovaných hodnot. Z grafu je i vidět, že preferovaný stav je zastoupen nejméně krát. Je to ale jen jedna z

možných realizací, takže se daný jev nedá zobecnit, ale i takového výsledku je možno dosáhnout. Je tedy jasné, že volit akce náhodně rovnoměrně určitě není správná cesta. Na druhou stranu četnosti stavů s optimalizací vyhověly preferenci agenta. Preferovaný stav je zastoupen skoro dvakrát více než ostatní stavy.

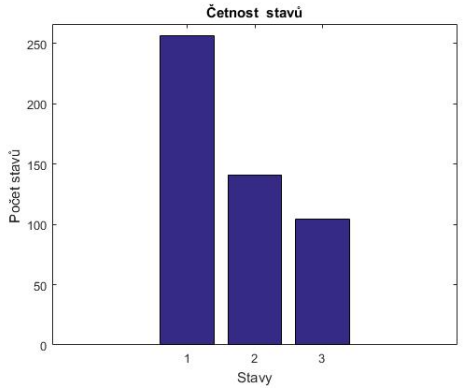
3.3 Ideál

Experiment 5. První experiment z druhé skupiny měl za úkol porovnávat vypočítané ideální optimální hodnoty π^i a p^i s předem pevně daným ideálem. Optimalizace i pevně daný ideál využíval bayesovské učení modelu systému. Pevně daný ideální model systému byl programátorsky označen "pi".

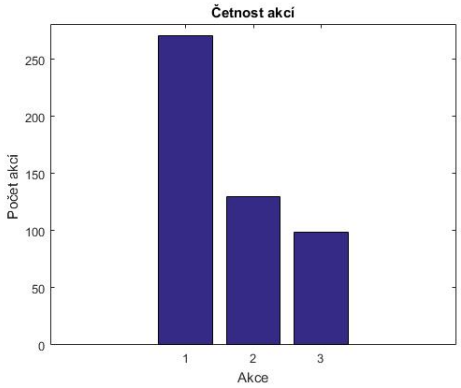
```
pi=zeros(|S|,|A|,|S|)
pi(1,1,1)=0.990  pi(2,1,1)=0.005  pi(3,1,1)=0.005
pi(1,2,1)=0.990  pi(2,2,1)=0.005  pi(3,2,1)=0.005
pi(1,3,1)=0.990  pi(2,3,1)=0.005  pi(3,3,1)=0.005

pi(:,:,3)=pi(:,:,2)=pi(:,:,1)
```

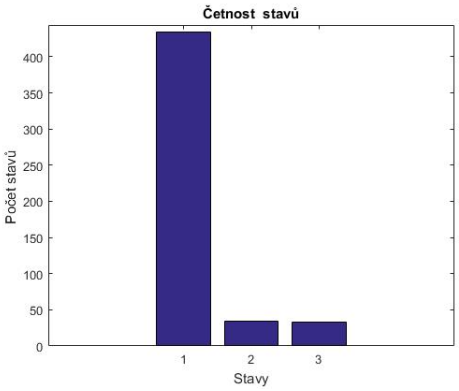
Obrázek 3.5: Experiment 5.



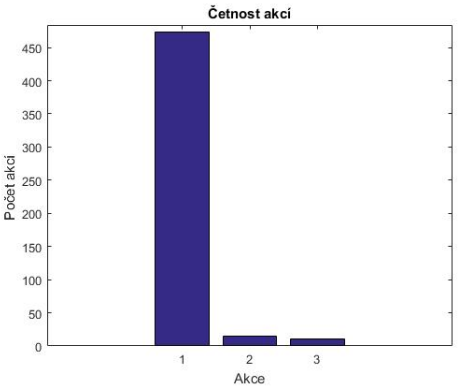
(1) Četnosti jednotlivých stavů s optimalizovaným ideálem a s učením, $s^i = 1$



(2) Četnosti jednotlivých akcí s optimalizovaným ideálem a s učením, $s^i = 1$



(3) Četnosti jednotlivých stavů s pevným ideálem a s učením, $s^i = 1$



(4) Četnosti jednotlivých akcí s pevným ideálem a s učením, $s^i = 1$

Tento ideál byl volen přímo pro nejpravděpodobnější hodnotu stavu $s^i = 1$. Jelikož nebyl žádný požadavek na akci, ideální rozhodovací pravidlo π^i bylo zvoleno rovnoměrně a bylo programátorsky označeno "pi_i".

```
pi_i=zeros(|A|,|S|)
pi_i(1,1)=0,333 pi_i(2,1)=0,333 pi_i(3,1)=0,333
pi_i(1,2)=0,333 pi_i(2,2)=0,333 pi_i(3,2)=0,333
pi_i(1,3)=0,333 pi_i(2,3)=0,333 pi_i(3,3)=0,333
```

Diskuze Tento experiment je jedním ze zásadních experimentů v této práci. Jak je patrno z grafů, pevný ideál dává lepší výsledky než optimalizovaný. Je tomu tak především kvůli jednoznačnosti preference a jednoduchosti systému. Pro větší systémy a pro více preferencí by bylo mnohem složitější vymyslet pevně daný ideál, který by splňoval dané preference. I když takto zvolený pevně daný ideál dává lepší výsledky než optimalizovaný ideál, je mnohem jednodušší a časově méně náročné jen zadat preferovaný stav a nechat optimalizaci vypočítat ideál, než nechat agenta pevně daný ideál vymyslet. Navíc agent pro využití optimalizace nemusí rozumět teorii rozhodovaní, jelikož to optimalizace vše vypočítá za něj.

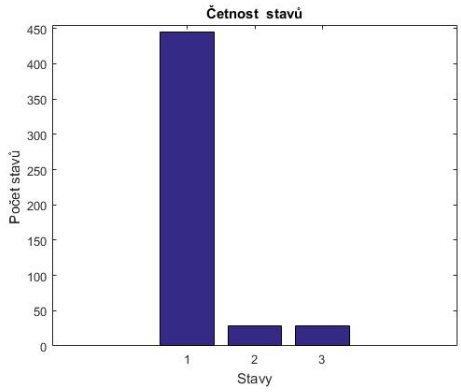
Experiment 6. Tento experiment opět porovnával vypočítané ideální optimální hodnoty π^i a p^i s předem pevně daným ideálem jako v předchozím experimentu, tentokrát byla však přidána preference na malé akce. Pevné ideální rozhodovací pravidlo bylo zvoleno

```
pi_i=zeros(|A|,|S|)
pi_i(1,1)=0,90 pi_i(2,1)=0,06 pi_i(3,1)=0,04
pi_i(1,2)=0,90 pi_i(2,2)=0,06 pi_i(3,2)=0,04
pi_i(1,3)=0,90 pi_i(2,3)=0,06 pi_i(3,3)=0,04
```

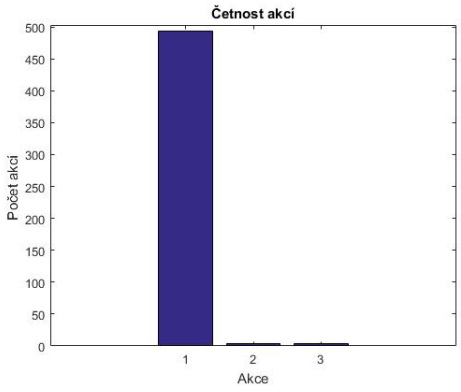
Pravděpodobnost akce $a = 1$ je nejvyšší a pravděpodobnost $a = 3$ je nejmenší. Pro vypočítané optimální hodnoty bylo zvoleno $\phi(a)$ tak, aby byly tyto dva experimenty porovnatelné.

```
phi(1)=1 phi(2)=5 phi(3)=6
```

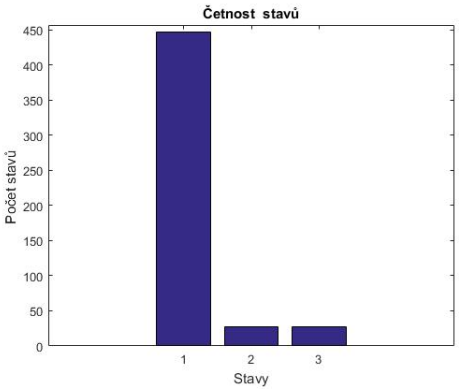
Obrázek 3.6: Experiment 6.



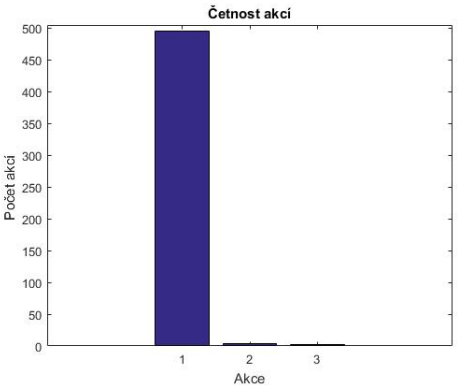
(1) Četnosti jednotlivých stavů s optimalizovaným ideálem s preferencí malých akcí a s učením, $s^i = 1$



(2) Četnosti jednotlivých akcí s optimalizovaným ideálem s preferencí malých akcí a s učením, $s^i = 1$



(3) Četnosti jednotlivých stavů s pevným ideálem s preferencí malých akcí a s učením, $s^i = 1$



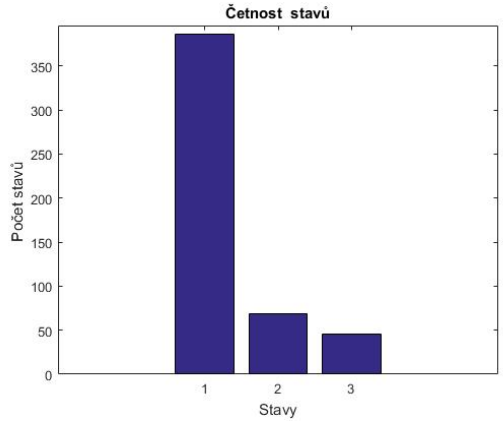
(4) Četnosti jednotlivých akcí s pevným ideálem s preferencí malých akcí a s učením, $s^i = 1$

Diskuze Jak je patrno z grafů, tak četnosti u optimalizovaného ideálu i pevně daného ideálu jsou velice podobné. Je tomu tak ale jen při volbě těchto dvou preferencí a naladění parametrů. Chtěla bych podotknout, že i když v tomto experimentu došlo téměř ke stejným výsledkům, je o mnoho příjemnější volit optimalizovaný ideál, kde se nastaví jen poměr akcí dle dané preference a poté se dá ještě pomocí parametru váhy $0 \leq \omega \leq 1$ ladit, jak moc jsou preferovány malé akce s ohledem na preferenci stavu.

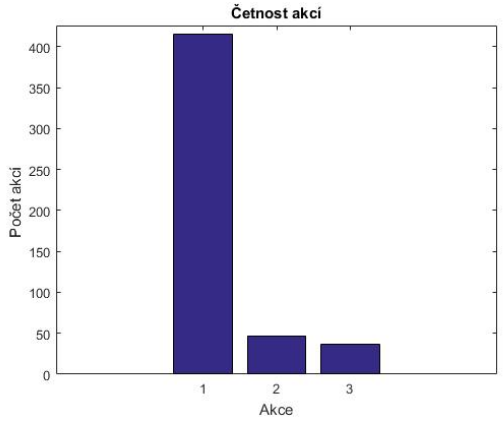
Kdyby byla zvolena ještě vyšší hodnota $\phi(3)$, byly by výsledky optimalizace ještě lepší. Pomocí ladění hodnot funkce $\phi(a)$ lze získat několik různých výsledků a je jen na agentovi, které hodnoty mu budou vyhovovat nejvíce. Využití optimálních hodnot ideálu je mnohem méně časově náročné. Při využití zpětné vazby agenta je očividně mnohem jednodušší změnit funkci $\phi(a)$, která má v tomto případě jen tři hodnoty, než změnit devět hodnot pravděpodobnosti ideálního rozhodovacího pravidla. Navíc při zadávání pevných hodnot je potřeba agentovo porozumění teorii rozhodovaní, což při zadání pouze preferovaných hodnot potřeba není.

Experiment 7. V tomto experimentu byly porovnány simulace s preferencí stavu $s^i = 1$ pro hodnoty váhy $\omega = 1$ a $\omega = 0$. Při volbě $\omega = 1$, by měly být zastoupeny malé hodnoty akcí častěji než s $\omega = 0$. Pro hodnotu váhy $\omega = 0$ splývá řešení s optimalizací výše.

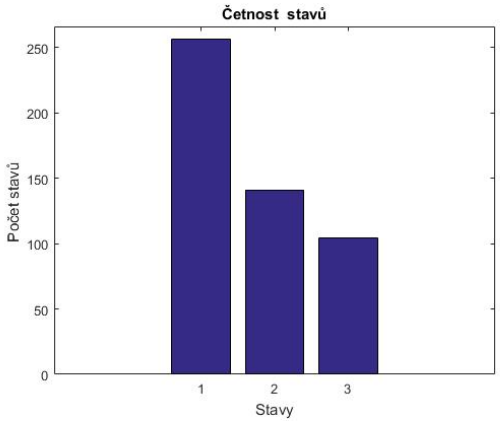
Obrázek 3.7: Experiment 7.



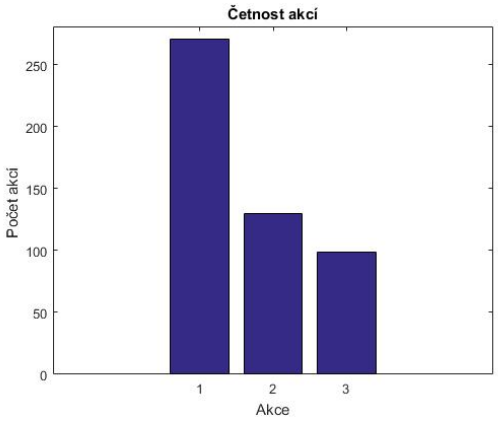
(1) Četnosti jednotlivých stavů s $\omega = 1$, $s^i = 1$



(2) Četnosti jednotlivých akcí s $\omega = 1$, $s^i = 1$



(3) Četnosti jednotlivých stavů s $\omega = 0$, $s^i = 1$



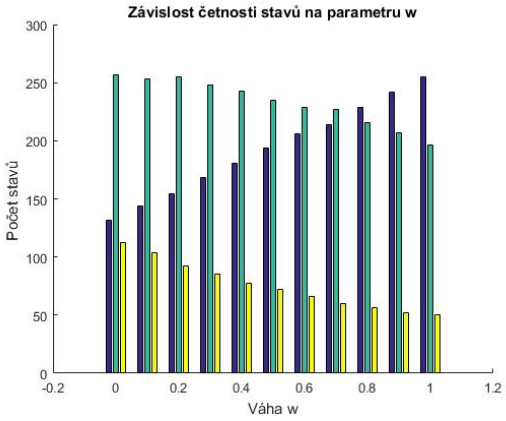
(4) Četnosti jednotlivých akcí s $\omega = 0$, $s^i = 1$

Diskuze Z grafů (Experiment 7.) je patrno, že pro hodnotu váhy $\omega = 1$ jsou mnohem četněji zastoupeny malé akce. Pro tento systém se preference $s^i = 1$ a malých akcí slučují, takže s váhou $\omega = 1$ jsou preference respektovány výrazně častěji a získají se lepší výsledky. Tyto dvě preference ale nejsou splněny vždy současně, proto se musí parametr ω naladit mezi 0 a 1 tak, aby bylo dosaženo požadovaných výsledků. Preference na akce nebude tak silná, pokud $0 \leq \omega \leq 1$ a povolí i odchylky od preference.

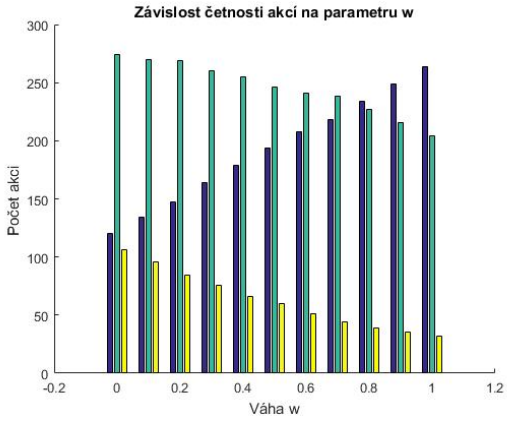
Pokud se ale tyto dvě preference navzájem nevylučují, jako v této úloze, četnost preferovaných stavů se zvýší.

Experiment 8. Osmá úloha navazuje na předchozí úlohu. Úkolem bylo vykreslení grafů závislosti četnosti stavů a akcí na hodnotě váhy ω, s preferencí stavu $s^i = 2$ a $s^i = 3$.

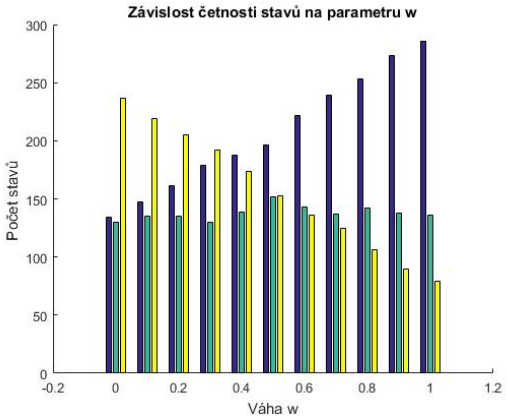
Obrázek 3.8: Experiment 8.



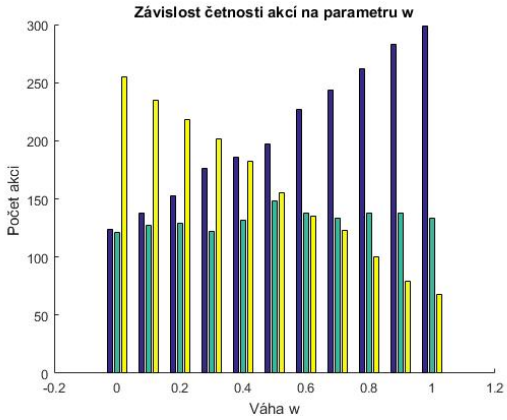
(1) Četnosti jednotlivých stavů v závislosti na ω, $s^i = 2$. Modrá pro stav 1, zelená stav 2, žlutá stav 3.



(2) Četnosti jednotlivých akcí v závislosti na ω, $s^i = 2$. Modrá pro akci 1, zelená akci 2, žlutá akci 3.



(3) Četnosti jednotlivých stavů v závislosti na ω, $s^i = 3$. Modrá pro stav 1, zelená stav 2, žlutá stav 3.

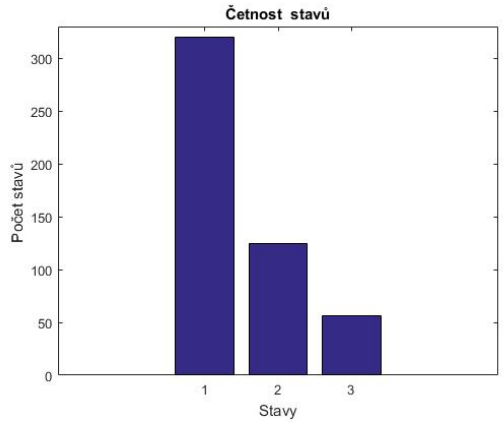


(4) Četnosti jednotlivých akcí v závislosti na w, $s^i = 3$. Modrá pro akci 1, zelená akci 2, žlutá akci 3.

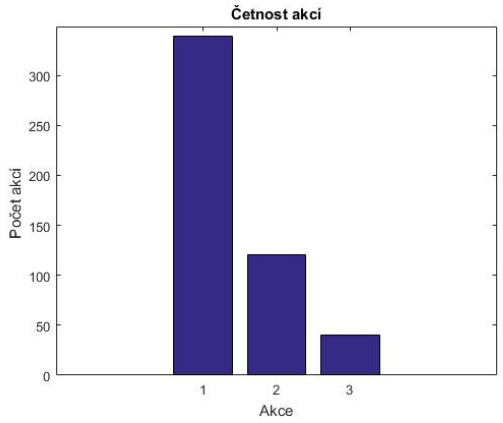
Diskuze Na tomto experimentu (Experiment 8.) je jasně vidět, že s rostoucím ω se zvyšuje četnost malých akcí ($a = 1$) a naopak četnost velkých akcí se snižuje ($a = 3$). Také je vidět, že preference stavu je dodržena pro $s^i = 2$ až hodnoty váhy $\omega = 0.7$, v této hodnotě se láme. Od této zlomové hodnoty je preference nízkých akcí silnější než preference stavu. Pro $s^i = 3$ je preference stavu dodržena pouze do hodnoty váhy $\omega = 0.3$, jelikož jsou tyto dvě preference ve větším nesouladu než pro případ $s^i = 2$.

Experiment 9. Tento experiment měl za úkol porovnávat četnosti stavů a akcí s váhou $\omega = 0.6$ s učením a bez učení.

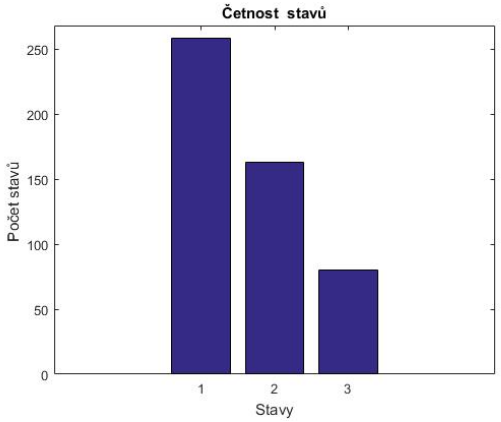
Obrázek 3.9: Experiment 9.



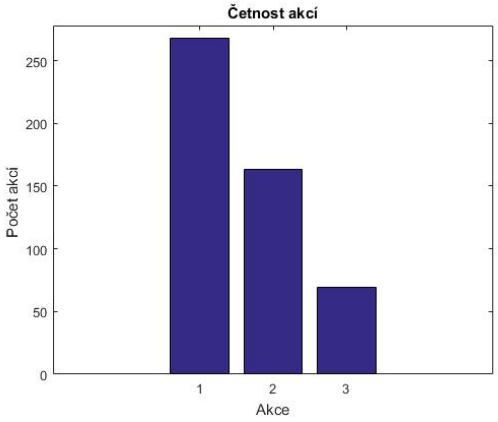
(1) Četnosti jednotlivých stavů s $\omega = 0.6$ s učením, $s^i = 1$



(2) Četnosti jednotlivých akcí s $\omega = 0.6$ s učením, $s^i = 1$



(3) Četnosti jednotlivých stavů s $\omega = 0.6$ bez učení, $s^i = 1$

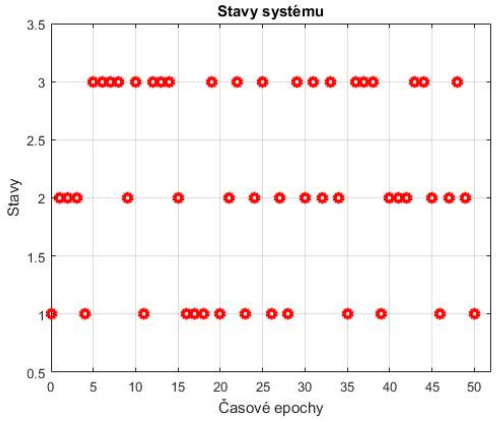


(4) Četnosti jednotlivých akcí s $\omega = 0.6$ bez učení, $s^i = 1$

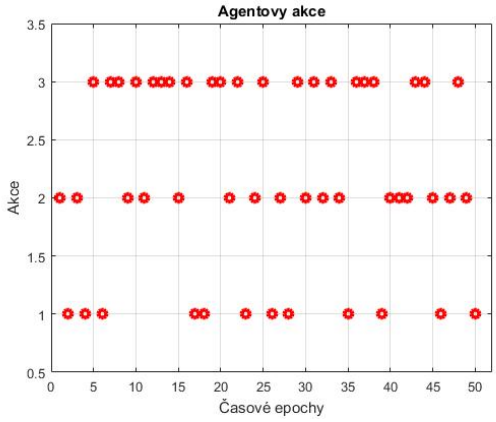
Diskuze V této úloze (Experiment 9.) je opět vidět, že optimalizace s učením dává mnohem lepší výsledky než bez, a to i pro preferenci malých akcí s váhou $\omega = 0.6$. Četnost stavů $s = 1$ je větší více než o padesát u optimalizace s učením než bez něj. A četnost dalších dvou stavů je samozřejmě výrazně nižší. Jak je vidět, samotná optimalizace také nestačí k dosažením nejlepších výsledků.

Experiment 10. Poslední experiment měl za úkol porovnávat posloupnosti stavů a akcí v jedné simulaci o délce simulace $d_{simulace} = 50$ (50 časových epoch) s váhou $\omega = 0$ a $\omega = 1$.

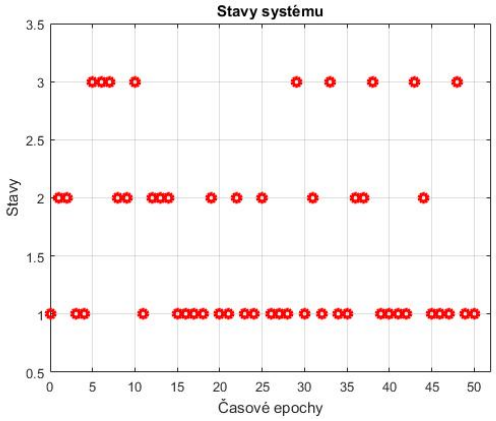
Obrázek 3.10: Experiment 10.



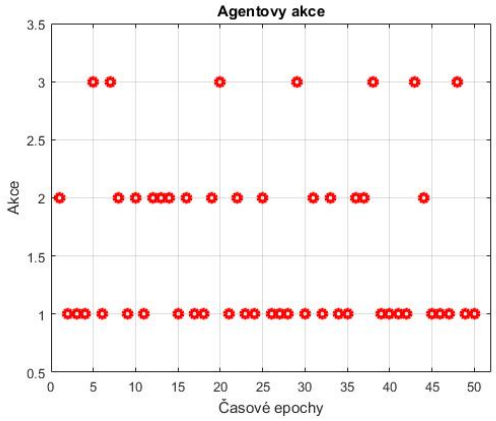
(1) Graf stavů v čase s $\omega = 0$, $s^i = 1$



(2) Graf akcí v čase s $\omega = 0$, $s^i = 1$



(3) Graf stavů v čase s $\omega = 1$, $s^i = 1$



(4) Graf akcí v čase s $\omega = 1$, $s^i = 1$

Diskuze Jak je vidět (Experiment 10.), četnosti preferovaných stavů jsou s váhou $\omega = 1$ vyšší než s $\omega = 0$. Tak tomu ale nebude vždy, jelikož se tyto dvě preference nebudou vždy v souladu. S touto preferencí $s^i = 1$ však preference malých akcí ještě více podpoří preferenci stavu.

Kapitola 4

Závěry

V této bakalářské práci jsem studovala teorii optimálního rozhodování na diskrétním markovském rozhodovacím procesu. Seznámila jsem se s plně pravděpodobnostním návrhem a dynamickým programováním. Naučila jsem se najít optimální rozhodovací pravidlo na základě daných preferencí pomocí **PPN**. Za pomoci článku [9] jsem se naučila hledat optimální ideální rozhodovací pravidlo a optimální ideální model systému.

Jelikož výše zmíněný článek pracoval se spojitým procesem, sestavila jsem obecnější rovnice pro nalezení optimálního ideálního modelu systému, které lze sestavit jak pro spojitý systém, tak pro diskrétní. A tyto rovnice mají řešení vždy. Dále jsem také přidala možnost preference na volbu akcí. Nakonec jsem vybrala několik experimentů, které ilustrují danou teorii a nasimulovala jsem je pomocí Matlabu. Zvolila jsem systém o třech stavech a třech akcích.

Jak jsem již zmiňovala v experimentální části, systém, který jsem pro tuto práci vybrala dává lepší výsledky než mohou dávat jiné systémy. Tento systém jsem vybírala záměrně tak, aby byla teorie lépe pochopena a aby výsledky co nejlépe ilustrovaly daný problém. Při zvolení jiných systémů nemusí optimalizace fungovat tak výrazně jako v tomto případě. Je jasné, že při zvolení systému, který má nízké pravděpodobnosti u stavu, který agent preferuje, nemusí dojít k nejvyšší četnosti preferovaného stavu, jen se dané četnosti preferovaného stavu zvýší oproti nepoužití žádné optimalizace.

Jak je také viditelné z experimentů, znalost systému výrazně napomáhá ve zlepšení optimalizace. Samotná optimalizace nestačí k dosažení nejlepších výsledků. Díky bayesovskému učení se vylepšuje model systému z již naměřených dat a agent tak získává určitou znalost systému, díky které je schopen s větší přesností určit, s jakou volbou akce se s největší pravděpodobností posune do preferovaného stavu. Úplná znalost systému je užita v Experimentu 1 a postupná částečná znalost byla získávána pomocí bayesovského učení například v Experimentech 2., 3.

To, že optimalizované rozhodovací pravidlo dává lepší výsledky než úplně náhodné rozhodovací pravidlo, je očividné z Experimentu 4. Proto hlavní informaci přinášejí experimenty zaměřené na optimální ideální model systému a na optimální rozhodovací pravidlo. Pomocí optimálního ideálu společně s bayesovským učením lze nalézt optimální politiku, která respektuje agentovy preference a získat tak velmi dobré výsledky. Jak ale ilustruje Experiment 5. i s pevně daným ideálem je možno dosáhnout dobrých výsledků. Chtěla bych ale zdůraznit výhody optimálního ideálu. Jeho použití je pro agenta velice pohodlné, je méně časově náročné a není k němu zapotřebí agentovo porozumění teorii rozhodování. Navíc při zadání neúplné preference či více preferencí není vůbec jednoduché zvolit pevný ideál, který by respektoval agentovy preference. Také při využití zpětné vazby agenta je časově náročné měnit hodnoty pevného ideálu.

Pomocí optimalizace lze však jen zvolit funkci $\phi(a)$, která vymezuje preferenci mezi jednotlivými akcemi, a nastavit parametr ω tak, aby byly současně splněny agentovy požadavky na stav. Pomocí

nastavení hodnoty parametru ω lze jednoduše vyjádřit vztah mezi preferencí stavu a akcí (nastavením parametru mezi 0 a 1 lze vyjádřit, jaká z preferencí je upřednostněna). Při využití zpětné vazby agenta je tak snadné a rychlé ladit parametry tak, aby výsledky odpovídaly co nejvíce agentovým požadavkům. To lze ale pomocí pevného ideálu velice obtížně. Agent musí plně rozumět teorii rozhodovaní a věnovat čas a vynaložit energii na zvolení tohoto ideálu, který navržená optimalizace "vymyslí" sama. Je dobré zdůraznit, že je opravdu velice náročné zvolit pevně dané ideály, pokud má agent více než jednu preferenci. Proto je nalezení optimálních ideálních hodnot tak přínosné.
Do budoucna bych se chtěla více věnovat složitějším systémům a praktikovat teorii na komplexnější problémy. Chtěla bych se zaměřit na větší systémy a řešit pomocí této optimalizace nějaký reálný případ.

Literatura

[1] R. Bellman. *Dynamic Programming*. Princeton Uni. Press, NY, 1957.

[2] J. Branke, S. Corrente, S. Greco, and W. Gutjahr. Efficient pairwise preference elicitation allowing for indifference. *Computers & Operations Research*, 88(Suppl. C):175 – 186, 2017.

[3] J. Drummond and C. Boutilier. Preference elicitation and interview minimization in stable matchings. In *Proceedins of Twenty-Eight AAAI Conference on AI*, pages 645 – 653, 2014.

[4] K. Gajos and D.S. Weld. Preference elicitation for interface optimization. In *Proc. of the 18th annual ACM Symposium on User interface and Technology, UIST-2005*, pages 173 – 182. 2005.

[5] I. Garcia, S. Pajares, L. Sebastia, and E. Onaindia. *Preference elicitation techniques for group recommender systems*, volume 189. Information Sciences, 2012.

[6] M. Kárný. Towards fully probabilistic control design. *Automatica*, 32(12):1719–1722, 1996.

[7] M. Kárný, J. Böhm, T.V. Guy, L. Jirsa, I. Nagy, P. Nedoma, and L. Tesař. *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. Springer, 2006.

[8] M. Kárný and T. V. Guy. Fully probabilistic control design. *Syst. & Con. Lett.*, 55(4):259–265, 2006.

[9] M. Kárný and T.V. Guy. Preference elicitation within framework of fully probabilistic design of decision strategies. In *IFAC International Workshop on Adaptive and Learning Control Systems - ALCOS 2019*, 2019. in print.

[10] M. Kárný and F. Hůla. Balancing exploitation and exploration via fully probabilistic design of decision policies. In *Proceedings of the 11th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART,*, pages 857–864. INSTICC, SciTePress, 2019.

[11] M. Kárný and T. Kroupa. Axiomatisation of fully probabilistic design. *Inf. Sci.*, 186(1):105–113, 2012.

[12] S. Kullback and R. Leibler. On information and sufficiency. *An. of Math. Stat.*, 22:79–87, 1951.

[13] A.M.D. Landmark, E.H. Ofstad, and J. Svenneig. *Eliciting patient preferences in shared decision-making (SDM): Comparing conversation analysis and SDM measurements*, volume 100. Patient Education and Counseling, 2017.

[14] L. Naamani-Dery, M. Kalech, L. Rokach, and B. Shapira. *Reducing preference elicitation in group decision making*, volume 61. Expert Systems with Applications, 2016.

[15] V. Peterka. Bayesian system identification. In P. Eykhoff, editor, *Trends and Progress in System Identification*, pages 239–304. Perg. Press, 1981.

[16] M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 2005.

[17] F. Rolenec. Feasible bayesian multi-armed bandit optimisation. Bachelor's thesis, FJFI, Czech Technical University, Prague, 2018.