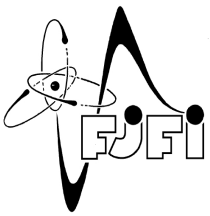
České vysoké učení technické v Praze
Fakulta jaderná a fyzikálně inženýrská



Plne pravdepodobnostný návrh rozhodovacej stratégie so zastavovaním

Fully probabilistic design of decision strategy with stopping

Bakalárska práca

Autor: **Mário Hoz**

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

Akademický rok: 2020/2021

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

Katedra: matematiky

Akademický rok: 2020/2021

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Student:	Mário Hoz
Studijní program:	Aplikace přírodních věd
Studijní obor:	Matematické inženýrství
Studijní zaměření:	Aplikované matematicko-stochastické metody
Název práce (česky):	Plně pravděpodobnostní návrh rozhodovací strategie se zastavováním
Název práce (anglicky):	Fully probabilistic design of decision strategy with stopping

Pokyny pro vypracování:

1) Seznamte se s plně pravděpodobnostním návrhem rozhodovacích strategií.
2) Seznamte se s formulací rozhodovacích problémů se zastavovacími pravidly.
3) Navrhněte optimální plně pravděpodobnostní strategii se zastavováním.
4) Algoritmizujte obecné řešení pro případ markovského systému s konečným počtem stavů a akcí.
5) Simulačně ověřte vlastnosti algoritmizovaného řešení.

Doporučená literatura:

1) M. Puterman, Markov decision processes. John Wiley & Sons, 1994.
2) M. Kárný, T. V. Guy, On the Origins of Imperfection and Apparent Non-Rationality. In 'T. V. Guy, M. Kárný, D. H. Wolpert, Decision Making: Uncertainty, Imperfection, Deliberation and Scalability', Springer, 2014, 57-92.
3) M. Kárný, T. V. Guy, Fully probabilistic control design. Systems & Control Letters, 55(4), 259-265, 2006.
4) V. Peterka, Bayesian System Identification. In. 'P. Eykhoff, Trends and Progress in System Identification', Pergamon Press, Oxford, 1981, 239-304.
5) A. Wald, Sequential Analysis. Dover Publications, 2013.

Jméno a pracoviště vedoucího bakalářské práce:

Ing. Miroslav Kárný, DrSc.
ÚTIA AV ČR, v.v.i., Pod vodárenskou věží 4, 18208 Praha 8

Jméno a pracoviště konzultanta:

Datum zadání bakalářské práce: 31.10.2020

Datum odevzdání bakalářské práce: 7.7.2021

Doba platnosti zadání je dva roky od data zadání.

V Praze dne 30.10.2020

.....
garant oboru

.....
vedoucí katedry

.....
děkan

Poďakovanie

Rád by som poďakoval svojmu školiteľovi Ing. Miroslavovi Kárnému, DrSc. za rýchlu komunikáciu, starostlivosť, ochotu a odborné aj ľudské zázemie pri vedení mojej bakalárskej práce.

Ďalej by som rád poďakoval oddeleniu adaptívnych systémov ÚTIA AV za možnosť podieľať sa na projektoch LTC18075 a EU-COST Action CA16228 podporovaných MŠMT ČR.

Mário Hoz

Prehlásenie

Prehlasujem, že som túto prácu vypracoval samostatne a uviedol som všetku použitú literatúru.

V Prahe dňa 16.7.2021

.....
Mário Hoz

Názov práce:
Plne pravdepodobnostný návrh rozhodovacej stratégie so zastavovaním

Autor: Mário Hoz

Program: Matematické inžinierstvo
Odbor: Aplikované matematicko-stochastické metódy
Druh práce: Bakalárska práca

Vedúci práce: Ing. Miroslav Kárný, DrSc.
ÚTIA AV ČR, Pod Vodárenskou věží 4, v.v.i. 182 08 Praha 8

Abstrakt:
Táto bakalárska práca sa zaoberá problematikou popisu zastavovania v teórií rozhodovacích procesov. Najprv sú popísané diskrétne Markove rozhodovacie procesy a ich zobecnenie v podobe plne pravdepodobnostného návrhu. Nasleduje detailné odvodenie vety pre nájdenie optimálnej politiky. Potom je diskutované rozšírenie plne pravdepodobnostného návrhu o popis zastavenia. Formuluje sa nielen tvar zastavovacích pravidiel, ale aj nových ideálov zohľadňujúcich zastavenie. V závere práce je veta pre nájdenie optimálnej politiky aplikovaná na takto rozšírený popis. Táto aplikácia odhaľuje neriešený problém, ktorý súvisí s návrhom ideálov zohľadňujúcich zastavenie.

Kľúčové slová: teória rozhodovania, Markov rozhodovací proces, plne pravdepodobnostný návrh, zastavovanie, zastavovacie pravidlo, dynamické programovanie

Title:
Fully probabilistic design of decision strategy with stopping

Author: Mário Hoz

Abstract:
This bachelor project deals with the description of stopping in the theory of decision processes. Firstly, discrete Markov decision processes and their generalization called the fully probabilistic design are described. Then, a theorem used for finding the optimal policy is derived in detail. Furthermore, a stopping extension of the fully probabilistic design is discussed, introducing not only the form of stopping rules, but also new ideals with stopping. Finally, the theorem used for finding the optimal policy is applied on the extended description. This application reveals an unsolved problem related to the design of ideals with stopping.

Key words: decision-making theory, Markov decision process, fully probabilistic design, stopping, stopping rule, dynamic programming

Obsah

Prehľad použitého značenia a skratiek

Značenie	Význam
$:=$	Definičné priradenie
$\mathcal{X} \sim P^{\mathcal{X}}$	Náhodná veličina $\mathcal{X}$ s rozdelením $P^{\mathcal{X}}$
$Ran\ \mathcal{X}$	Obor hodnôt $\mathcal{X}$
$f(x)$	Hustota pravdepodobnosti $\mathcal{X}$
$f(x_2 \mid x_1)$	Hustota pravdepodobnosti $\mathcal{X}_2$ podmienená vzťahom $\mathcal{X}_1 = x_1$
$\mathbb{T}, \mathbb{A}, \mathbb{S}, \mathbb{B}$	Množiny
$\mathbb{N}$	Množina všetkých prirodzených čísel
$\mathbb{R}$	Množina všetkých reálnych čísel
Π	Množina rozhodovacích pravidiel
$\{x^i\}_{i=1}^m$	Množina prvkov $x^1, x^2, \ldots, x^n$
$(x_t)_{t=1}^H$	Postupnosť prvkov $x_1, x_2, \ldots, x_H$
$\min_{x \in \mathbb{R}} f(x)$	Minimum funkcie $f(x)$ na množine $\mathbb{R}$
$\arg[\cdot]$	Hodnota argumentu
$\mathrm{E}[\cdot]$	Očakávaná hodnota
$\mathrm{E}[\cdot \mid \cdot]$	Podmienená očakávaná hodnota
$D\left(\cdot \| \cdot\right)$	Kullback-Leiblerova divergencia

Skratka	Význam
KLD	Kullback-Leiblerova divergencia
MRP	Markov rozhodovací proces
PPN	Plne pravdepodobnostný návrh

Úvod

Každý z nás musí denne urobiť stovky rozhodnutí. V pozícií rozhodovacieho agenta vstupujeme do situácií, v ktorých sa musíme rozhodnúť na základe známeho a pozorovaného v danom systéme. Pozorujeme systém v istom stave a reagujeme jednou z možných akcií. Akcia spätne ovplyvní systém, čím je generovaný nový stav a situácia sa opakuje. Našou motiváciou je samozrejme dosiahnuť vytúženého cieľa, resp. vyhnúť sa nepriaznivému koncu.

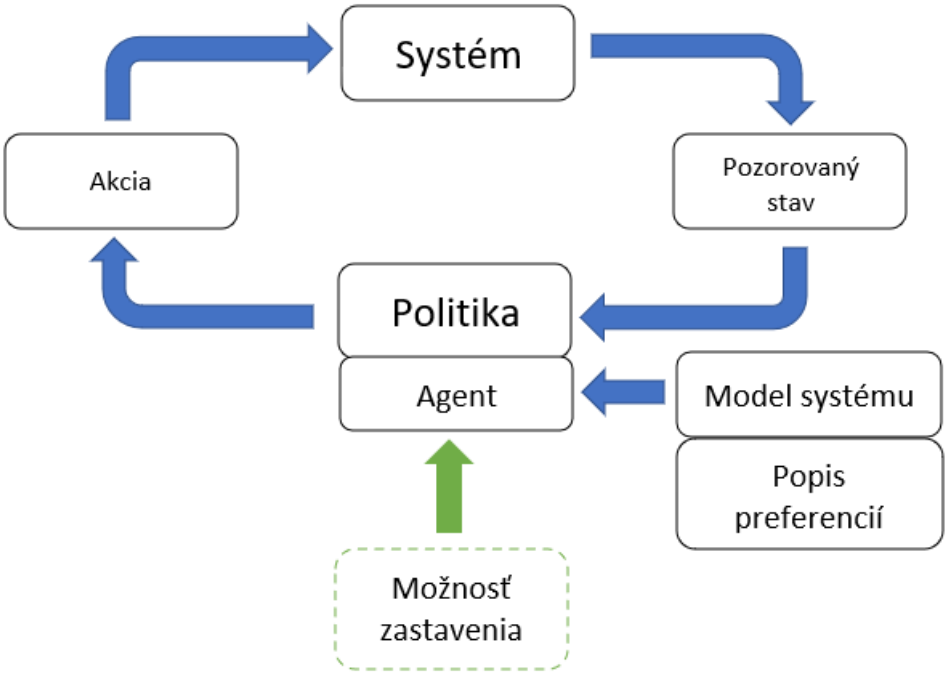
Súbor pravidiel, na základe ktorých akcie vyberáme, sa nazýva politika. Politika tvorí so systémom uzavretú slučku. Pre základný matematický popis celej uzavretej slučky sa používajú najmä Markove rozhodovacie procesy. Tie sú detailnejšie popísané v [1] a v podkapitole 1.2 tejto práce. Na Markove rozhodovacie procesy budeme odkazovať skratkou MRP. K tvorbe politiky agent užíva model systému a preferencie (svoje vlastné, aj dodané z okolia). Popis modelu systému, ako aj popis preferencií sú do uzavretej slučky na začiatku externe vložené.

Preferencie vstupujú do návrhu tzv. optimálnej politiky, ktorá popisuje optimálny scenár k dosiahnutiu cieľa. Je dôležité preferencie adekvátne kvantifikovať a zaistiť ich konzistentnosť (viz [2]).

Model systému získavame napr. z odborovej znalosti. Obecne je možné sa model systému aj postupne učiť. Dáta v podobe postupnosti pozorovaných stavov a akcií sa dajú využiť v procese učenia na opravu modelu a následné zlepšenie zvolenej politiky. Neznáme parametre modelu možno bayesovsky odhadovať (viz [3]).

V praxi je získavanie väčšieho množstva informácií o systéme doprevádzané vynaložením prostriedkov. Je nutné investovať čas, peniaze, či iné zdroje agenta. Za každé pozorovanie systému s cieľom získať novú informáciu platí agent istú čiastku. Táto čiastka môže byť fixná, alebo sa môže priebežne meniť podľa okolností. Náklady sa každopádne kumulujú.

Z vyššie uvedeného plynie kritická otázka, ktorej každý agent čelí. Ide o určenie momentu, kedy prestať zhromažďovať nové informácie. Tento moment môže predchádzať momentu ukončenia celého procesu, tzn. agent môže len zastaviť učenie a v samotnom rozhodovacom procese pokračovať ďalej.



Obr. 1: Uzavretá slučka pre proces rozhodovania.
Zelený blok znázorňuje časť, ktorej sa táto práca primárne venuje.

Spomenutá problematika zastavovania je veľmi rozsiahla.

V histórií možno nájsť mnoho optimalizačných úloh, ktoré sa zaoberajú hľadaním konkrétneho zastavovacieho pravidla. Toto hľadanie vychádza zo sekvenčnej analýzy popísanej v [4]. Výsledné zastavovacie pravidlo môže mať dokonca podobu kombinácie viacerých zastavovacích pravidiel, medzi ktorými agent operatívne prepína. Takéto kombinovanie v rámci *teórie výberu zastavovacieho pravidla* naznačuje podľa [5] sľubné výsledky.

Jedným zo známych príkladov skúmania zastavovacieho pravidla je tzv. problém sekretárky. Ten dal vzniknúť celej novej oblasti v rámci optimalizačných úloh. Zatiaľ čo [6] popisuje históriu tohto problému, jeho kompletné riešenie je podrobne rozpísané v [7]. Napriek detailnému preskúmaniu problému sekretárky podľa [8] účastníci experimentu nie sú vždy schopní konať na základe optimálneho zastavovacieho pravidla.

Práca [5] diskutuje, že zanedbanie optimálnych zastavovacích pravidiel sa nevyskytuje len u problému sekretárky, ale je bežným javom. Existujúce optimalizácie často nereflektujú reálne chovanie agenta a nezohľadňujú jeho prípadnú iracionalitu, ktorá spôsobuje, že niektoré časti uzavretej slučky sú pri rozhodovaní zanedbané. Agent má tendenciu zastaviť zber dát príliš skoro, alebo vykonať také rozhodnutie, pre ktoré má nedostatočnú evidenciu.

Ukazuje sa, že optimálny model možno aplikovať len vtedy, keď sa agent pohybuje v tzv. malom svete (viz [9]). Malý svet je označenie pre takú úlohu, o ktorej má agent plnú vedomosť a všetky premenné sú mu známe.

Väčšinou však uvažujeme pohyb agenta vo veľkom svete, ktorý je veľmi nepredvídateľný a dynamický, s množstvom neznámych premenných. Z každodennej skúsenosti vieme, že naše rozhodovanie ovplyvňuje aj náš aktuálny psychický stav, ktorý nie je matematicky dobre uchopiteľný. Pôvod nedokonalosti a iracionálnosti agenta, ako aj následná náprava sú jedným z významných predmetov súčasného výskumu v teórií rozhodovania. O tejto problematike viac pojednávajú napr. [10], [11] a [12]. Prípady, kedy teória nepodporuje reálne rozhodovanie agenta, obecne motivujú vznik tejto práce.

Jej cieľom je obecný (abstraktný) popis optimalizácie zastavovacích pravidiel. Pre tento obecný popis optimalizácie využijeme teóriu plne pravdepodobnostného návrhu, ktorá je axiomaticky zavedená v [13]. Plne pravdepodobnostný návrh, na ktorý budeme odkazovať skratkou PPN, umožňuje efektívne popisovať dynamické rozhodovanie a pracovať s náhodilosťou v chovaní agenta. PPN však nemá dostatočne rozpracovanú problematiku rozhodovania so zastavovaním. Hlavnou motiváciou tejto práce je snaha o nápravu tohto stavu.

V prvej kapitole práce sú uvedené kľúčové definície a vety z teórie pravdepodobnosti, o ktoré sa v tejto práci opierame. Ďalej je stručne popísané fungovanie MRP a zavedený PPN, ktorý z MRP vychádza. Hlavným rozdielom medzi MRP a PPN je spôsob vyjadrenia strát, ktoré sú utŕžené počas rozhodovacieho procesu.

V druhej kapitole je predstavený problém zastavovania. Práca diskutuje zatiaľ nedostatočne riešenú otázku popisu zastavovacích pravidiel v PPN. Zastavovacie pravidlá sú zavedené pomocou akcie predĺženia. Tá určuje, či má daný proces pokračovať (byť predĺžený), alebo sa zastaviť. Pre jednoduchšie rozlíšenie potom nazývame akciu pôvodnej úlohy ako normálna akcia. Jadro práce tvorí snaha o zobecnenie vety o zastavovaní z [10]. Toto zobecnenie by malo popisovať návrh optimalizácie pre ľubovoľné zastavovacie pravidlo pomocou rozšíreného dynamického programovania v PPN. V poslednej časti kapitoly je uvedený hlavný problém, na ktorý sa počas snahy o toto zobecnenie narazilo.

Záver sumarizuje najdôležitejšie výsledky tejto práce a diskutuje otvorené problémy pre ďalšie skúmanie.

Kapitola 1

Matematická formulácia

1.1 Teória pravdepodobnosti

V tejto kapitole uvádzame len vybrané definície a vety, ktoré sú relevantné pre zavedenie Markových rozhodovacích procesov a plne pravdepodobnostného návrhu. Teória pravdepodobnosti je rozsiahlo popísaná napr. v [14] a [15], kde možno nájsť aj dôkazy nižšie uvedených viet.

V celej práci uvažujeme diskrétne náhodné veličiny, ak nie je explicitne uvedené inak.

Definícia 1.1.1 (Diskrétna náhodná veličina)
Označme $\mathcal{X} \sim P^{\mathcal{X}}$ náhodnú veličinu $\mathcal{X}$ s pravdepodobnostným rozdelením $P^{\mathcal{X}}$.
Túto náhodnú veličinu nazveme **diskrétna** práve vtedy, keď jej obor hodnôt $\operatorname{Ran}\mathcal{X}$ je najviac spočetný, tzn. $\operatorname{Ran}\mathcal{X} = \{x^i\}_{i=1}^{m}$, kde $m \in \mathbb{N} \cup \{+\infty\}$.

Definícia 1.1.2 (Hustota pravdepodobnosti)
Nech $\mathcal{X} \sim P^{\mathcal{X}}$ diskrétna náhodná veličina, $\operatorname{Ran}\mathcal{X} = \{x^i\}_{i=1}^{m}$.
Zaveďme pravdepodobnosť $p^i := P(\mathcal{X} = x^i)$.
Hustota pravdepodobnosti $f(x)$ je definovaná pre $\forall x \in \mathbb{R}$ vzťahom

$$f(x) := \begin{cases} p^i & \text{ak } x = x^i \in \operatorname{Ran}\mathcal{X}, \\ 0 & \text{inak.} \end{cases}$$

Náhodnú veličinu, ktorej prísluší daná hustota pravdepodobnosti, identifikujeme v tejto práci výhradne podľa argumentu hustoty. Keďže pravdepodobnostná miera je z Kolmogorových axiómov nezáporná a normovaná na jednotku, platí $f(x) \geq 0, \sum_{x \in \mathbb{R}} f(x) = 1$.

Definícia 1.1.3 (Marginálna hustota pravdepodobnosti)
Nech $\mathcal{X} \sim P^{\mathcal{X}}$ diskrétna n-rozmerná náhodná veličina.
Pre ľubovoľné $j \in \{1, \ldots, n\}$ uvažujme $(n-1)$-rozmernú náhodnú veličinu $\mathcal{X}' \sim P^{\mathcal{X}'}$, ktorá vzniká z $\mathcal{X}$ vynechaním j-tej zložky.
Marginálna hustota pravdepodobnosti $f(x')$ je definovaná pre $\forall x' \in \mathbb{R}^{n-1}$ vzťahom

$$f(x') := \sum_{x_j} f(x).$$

Obdobne možno získať vysčítaním cez viaceré premenné marginálne hustoty pravdepodobnosti pre náhodné veličiny s menším počtom zložiek. V tejto práci budeme najčastejšie uvažovať trojrozmerné náhodné veličiny $\mathcal{X} = (\mathcal{X}_3, \mathcal{X}_2, \mathcal{X}_1)$ a marginálne hustoty pravdepodobnosti

$$f(x_1) = \sum_{x_3, x_2} f(x_3, x_2, x_1).$$

Definícia 1.1.4 (Podmienená hustota pravdepodobnosti)
Nech $\mathcal{X} = (\mathcal{X}_2, \mathcal{X}_1) \sim f(x_2, x_1)$ diskrétna náhodná veličina.
Podmienená hustota pravdepodobnosti $f(x_2 \mid x_1)$ za podmienky $\mathcal{X}_1 = x_1$ je definovaná pre $\forall x_2 \in \operatorname{Ran} \mathcal{X}_2$ a $\forall x_1 \in \operatorname{Ran} \mathcal{X}_1$ kde $f(x_1) \neq 0$ vzťahom

$$f(x_2 \mid x_1) := \frac{f(x_2, x_1)}{f(x_1)}.$$

Veta 1.1.1 (Reťazové pravidlo)
Nech $\mathcal{X} = (\mathcal{X}_2, \mathcal{X}_1) \sim f(x_2, x_1)$ diskrétna náhodná veličina.
Z definície podmienenej hustoty pravdepodobnosti máme

$$f(x_2, x_1) = f(x_2 \mid x_1) f(x_1) = f(x_1 \mid x_2) f(x_2).$$

Obecne pre $\mathcal{X} = (\mathcal{X}_n, \mathcal{X}_{n-1}, \ldots, \mathcal{X}_1)$ platí **reťazové pravidlo** v tvare

$$\begin{aligned} f(x) &= f(x_n \mid x_{n-1}, \ldots, x_1) f(x_{n-1}, \ldots, x_1) = \\ &= f(x_n \mid x_{n-1}, \ldots, x_1) f(x_{n-1} \mid x_{n-2}, \ldots, x_1) \ldots f(x_2 \mid x_1) f(x_1) = \\ &= f(x_1) \prod_{k=2}^{n} f(x_k \mid x_{k-1}, \ldots, x_1). \end{aligned} \tag{1.1}$$

Veta 1.1.2 (Bayesovo pravidlo)
Nech $\mathcal{X} = (\mathcal{X}_2, \mathcal{X}_1) \sim f(x_2, x_1)$ diskrétna náhodná veličina.
Pre $\forall x_2 \in \operatorname{Ran} \mathcal{X}_2$ a $\forall x_1 \in \operatorname{Ran} \mathcal{X}_1$ kde $f(x_1) \neq 0$ platí **Bayesovo pravidlo** v tvare

$$f(x_2 \mid x_1) = \frac{f(x_1 \mid x_2) f(x_2)}{f(x_1)}. \tag{1.2}$$

Menovateľ možno vyjadriť z definície marginálnej a podmienenej hustoty pravdepodobnosti ako

$$f(x_1) = \sum_{x_2} f(x_2, x_1) = \sum_{x_2} f(x_1 \mid x_2) f(x_2).$$

1.2 Diskrétny Markov rozhodovací proces

Základným matematickým popisom pre sekvenciu udalostí je diskrétny Markov reťazec.

Riadený Markov reťazec, do ktorého vstupujú aj akcie agenta, nazývame Markov rozhodovací proces. V tejto práci ho označujeme skratkou MRP.

Definícia 1.2.1 (Diskrétny Markov rozhodovací proces)
Diskrétny Markov rozhodovací proces je usporiadaná pätica $(\mathbb{T}, \mathbb{S}, \mathbb{A}, p, l)$.

- **Množina časov** $\mathbb{T} = \{1, \ldots, H\}$, kde $H \in \mathbb{N}$ je pevné.
  Číslo H nazývame **konečný horizont**.
- **Množina stavov** $\mathbb{S} = \{s^i\}_{i=1}^m$, kde $m \in \mathbb{N}$ je pevné.
  Táto množina vzniká zjednotením možných stavov pre všetky uvažované časy.
  Jej prvky sú pevne dané.
- **Množina akcií** $\mathbb{A} = \{a^i\}_{i=1}^n$, kde $n \in \mathbb{N}$ je pevné.
  Táto množina vzniká zjednotením dostupných akcií pre všetky uvažované časy.
  Jej prvky sú pevne dané. Pre čas $t \in \mathbb{T}$ uvažujeme akciu $a_t = a \in \mathbb{A}$, ktorá uvedie systém zo stavu $s_{t-1} = s \in \mathbb{S}$ do stavu $s_t = s' \in \mathbb{S}$. [1]
- **Prechodová funkcia** $p(s_t \mid a_t, s_{t-1})$
  Podmienená hustota pravdepodobnosti popisujúca prechod zo stavu $s_{t-1} = s \in \mathbb{S}$ vykonaním akcie $a_t = a \in \mathbb{A}$ práve do stavu $s_t = s' \in \mathbb{S}$. Počiatočný stav systému s_0 je známy a pevný. Tento počiatočný stav je implicitnou súčasťou všetkých podmienok.
- **Stratová funkcia** $l(s_t, a_t, s_{t-1})$
  Reálna funkcia popisujúca **stratu** $l(s_t = s', a_t = a, s_{t-1} = s) \in \mathbb{R}$, ktorú agent obdrží po vykonaní akcie $a_t = a \in \mathbb{A}$, ktorou uvedie systém zo stavu $s_{t-1} = s \in \mathbb{S}$ práve do stavu $s_t = s' \in \mathbb{S}$.

Definícia 1.2.2 (Rozhodovacie pravidlo)
Rozhodovacie pravidlo $\pi(a_t \mid s_{t-1})$ je podmienená hustota pravdepodobnosti, ktorá popisuje výber akcie $a_t = a \in \mathbb{A}$ pre stav $s_{t-1} = s \in \mathbb{S}$.

Definícia 1.2.3 (Politika)
Politika π je postupnosť rozhodovacích pravidiel $\pi := (\pi(a_t \mid s_{t-1}))_{t=1}^H$.
Množinu všetkých uvažovaných politík označíme ako Π.

V každom čase agent pozoruje istý stav systému a užíva isté rozhodovacie pravidlo. Na základe rozhodovacieho pravidla v danom stave agent vyberie akciu, ktorá má dva dôsledky. Za prvé, agent obdrží okamžitú stratu popísanú stratovou funkciou. Za druhé, systém prejde do nového stavu s pravdepodobnosťou popísanou pomocou prechodovej funkcie. Každý nový stav sa vlastne chová ako (akciami ovplyvnená) realizácia náhodnej veličiny.

[1] Využívame konvenciu, podľa ktorej stotožňujeme index akcie s indexom stavu *do* ktorého sa touto akciou dostávame. Oneskorenie v rámci uzavretej slučky tak pripisujeme agentovi (politike).

V rámci tejto práce uvažujeme výhradne uzavreté slučky, v ktorých systém aj politika (rozhodovacie pravidlá) majú Markovu vlastnosť. Markova vlastnosť znamená, že podmienené hustoty pravdepodobnosti, ktoré systém a politiku popisujú, závisia bezprostredne na predchádzajúcom stave a nie na celej histórií. Systém a politika s Markovou vlastnosťou fungujú „len s pamäťou 1“.

Obecným cieľom rozhodovania sú čo najmenšie celkové straty

$$L := \sum_{t \in \mathbb{T}} l(s_t, a_t, s_{t-1}). \tag{1.3}$$

Aby bola takáto minimalizácia možná, potrebovali by sme poznať budúce stavy. My o nich ale znalosť nemáme. Preto v každom čase minimalizujeme celkové očakávané straty.

Definícia 1.2.4 (Očakávané straty)
Čiastkové očakávané straty sú v každom čase $t \in \mathbb{T}$ definované ako

$$\mathrm{E}^{\pi}[l(s_t, a_t, s_{t-1})] := \sum_{\substack{s_t, s_{t-1} \in \mathbb{S} \\ a_t \in \mathbb{A}}} l(s_t, a_t, s_{t-1}) p(s_t \mid a_t, s_{t-1}) \pi(a_t \mid s_{t-1}) p(s_{t-1}). \tag{1.4}$$

Celkové očakávané straty definujeme ako

$$\mathrm{E}^{\pi}[L] := \sum_{t \in \mathbb{T}} \sum_{\substack{s_t, s_{t-1} \in \mathbb{S} \\ a_t \in \mathbb{A}}} l(s_t, a_t, s_{t-1}) p(s_t \mid a_t, s_{t-1}) \pi(a_t \mid s_{t-1}) p(s_{t-1}). \tag{1.5}$$

Užívame dohody $\mathrm{E}[\cdot] = \mathrm{E}[\cdot \mid s_0]$. Horným indexovaním podľa π vyjadrujeme, že očakávané straty sú priamo závislé na politike. V MRP politika vstupuje do očakávaných strát lineárne.

Ak je politika deterministická, jej aplikovaním sa stáva z Markovho rozhodovacieho procesu len špecifický Markov reťazec. Na základe rozhodovacích pravidiel daných touto politikou vyberáme v každom časovom okamihu pre $s_{t-1} = s \in \mathbb{S}$ jednoznačne určenú akciu $a_t = a(s_{t-1}) \in \mathbb{A}$. Tvar očakávaných strát sa potom pre deterministickú politiku zjednoduší na

$$\mathrm{E}^{\pi}[l(s_t, a_t, s_{t-1})] = \sum_{s_t, s_{t-1} \in \mathbb{S}} l(s_t, a_t(s_{t-1}), s_{t-1}) p(s_t \mid a_t(s_{t-1}), s_{t-1}) p(s_{t-1}),$$
$$\mathrm{E}^{\pi}[L] = \sum_{t \in \mathbb{T}} \sum_{s_t, s_{t-1} \in \mathbb{S}} l(s_t, a_t(s_{t-1}), s_{t-1}) p(s_t \mid a_t(s_{t-1}), s_{t-1}) p(s_{t-1}).$$

Ďalej budeme uvažovať podmienené čiastkové očakávané straty pri deterministickej politike. Pri predpoklade známeho $s_{t-1} = s \in \mathbb{S}$ dostaneme

$$\mathrm{E}^{\pi}[l(s_t, a_t, s_{t-1}) \mid s_{t-1} = s] = \sum_{s_t \in \mathbb{S}} l(s_t, a_t(s_{t-1}), s_{t-1}) p(s_t \mid a_t(s_{t-1}), s_{t-1}).$$

Pre minimalizáciu tohto výrazu platí

$$\min_{\pi \in \Pi} \mathrm{E}^{\pi}[l(s_t, a_t, s_{t-1}) \mid s_{t-1} = s] = \min_{a \in \mathbb{A}} \mathrm{E}[l(s_t, a_t(s_{t-1}), s_{t-1}) \mid a_t(s_{t-1}) = a, s_{t-1} = s].$$

Definícia 1.2.5 (Hodnotová funkcia)
Uvažujme MRP $(\mathbb{T}, \mathbb{S}, \mathbb{A}, p, l)$, čas $t \in \mathbb{T}$ a politiku π.
Hodnotová funkcia politiky π je funkcia $u^\pi : \mathbb{S} \to \mathbb{R}$ vyjadrujúca budúcu kumuláciu očakávaných strát pri rozhodovaní podľa π za podmienky, že vychádzame zo stavu $s_{t-1} = s \in \mathbb{S}$. Je definovaná ako

$$u^\pi(s_{t-1}) := \mathrm{E}^\pi \left[\sum_{\substack{\tau \in \mathbb{T} \\ \tau \geq t}} l(s_\tau, a_\tau, s_{\tau-1}) \mid s_{t-1} = s \right] \text{ pre } \forall s \in \mathbb{S}. \tag{1.6}$$

V čase $t = H$ je hodnotová funkcia dodefinovaná nulou, $u^\pi(s_H) := 0$ pre $\forall s \in \mathbb{S}$.

Definícia 1.2.6 (Optimálna hodnotová funkcia)
Uvažujme MRP $(\mathbb{T}, \mathbb{S}, \mathbb{A}, p, l)$, čas $t \in \mathbb{T}$ a politiku π.
Optimálna hodnotová funkcia je definovaná ako hodnotová funkcia splňujúca

$$u^{\pi^*}(s_{t-1}) := \min_{\substack{\{\pi(a_\tau | s_{\tau-1})\} \\ \tau \geq t}} \mathrm{E}^\pi \left[\sum_{\substack{\tau \in \mathbb{T} \\ \tau \geq t}} l(s_\tau, a_\tau, s_{\tau-1}) \mid s_{t-1} = s \right] \text{ pre } \forall s \in \mathbb{S}. \tag{1.7}$$

Pre čas $t = 1$ dostávame $u^{\pi^*}(s_0) = \min_{\pi \in \Pi} \mathrm{E}^\pi[L] = \min_{\pi \in \Pi} u^\pi(s_0)$, kde s_0 je známy počiatočný stav. Optimálna hodnotová funkcia v tomto prípade splýva s minimalizáciou celkových očakávaných strát cez všetky politiky. Preto možno hodnotovú funkciu využiť na zmeranie kvality istej politiky, resp. na nájdenie optimálnej politiky.

Definícia 1.2.7 (Optimálna politika)
Uvažujme MRP $(\mathbb{T}, \mathbb{S}, \mathbb{A}, p, l)$. **Optimálnou politikou** π^* rozumieme politiku minimalizujúcu celkové očakávané straty. Pomocou hodnotovej funkcie ju možno vyjadriť ako

$$\pi^* \in \arg\min_{\pi \in \Pi} u^\pi(s_0). \tag{1.8}$$

Optimálna politika nemusí byť jednoznačná, preto ju uvažujeme ako prvok príslušnej množiny.

Pri hľadaní optimálnej politiky čelíme problému, že voľba akcií musí byť uskutočnená bez znalosti budúcich stavov. To robí z nájdenia optimálnej politiky zložitú úlohu. Spôsob, akým možno optimálnu politiku nájsť, sa nazýva dynamické programovanie (viz [16]). Pri dynamickom programovaní určíme koncový stav, do ktorého sa chceme dostať. Z tohto stavu spätným chodom optimalizujeme jednotlivé kroky.

Veta 1.2.1 (Dynamické programovanie)
Uvažujme MRP $(\mathbb{T}, \mathbb{S}, \mathbb{A}, p, l)$. Potom optimálnu hodnotovú funkciu možno vypočítať pre $\forall t \in \mathbb{T}$ spätnou rekurziou začínajúcou v $t = H$ pomocou vzťahu

$$u^{\pi^*}(s_{t-1}) = \min_{a \in \mathbb{A}} \mathrm{E}^{\pi} \left[l(s_t, a_t, s_{t-1}) + u^{\pi^*}(s_t) \mid a_t = a, s_{t-1} = s \right]. \tag{1.9}$$

V každom čase vyberáme za optimálne také rozhodovacie pravidlo, ktoré generuje tú akciu $a_t = a \in \mathbb{A}$, pre ktorú je minimum (1.9) dosahované.

Dôkaz. Je len špeciálnym prípadom vety 1.4.4 dokázanej v kapitole 1.4. □

Dynamické programovanie prakticky funguje podľa nasledujúcej schémy:

- Uvažujeme $t = H$ a náš koncový stav, do ktorého sa chceme dostať.
- Keďže pre $\forall s_H = s' \in \mathbb{S}$ je $u^{\pi^*}(s_H) = 0$, v prvom kroku počítame pre $\forall s_{H-1} = s \in \mathbb{S}$ len minimalizáciu čiastkových očakávaných strát.
- Pre každý stav $s \in \mathbb{S}$ určíme rozhodovacie pravidlo, ktoré vyberá minimalizujúcu akciu

$$\arg\min_{a \in \mathbb{A}} \mathrm{E}^{\pi}[l(s_H, a_H, s_{H-1}) \mid a_H = a, s_{H-1} = s].$$

- Keďže $\mathbb{A}$ je neprázdna a konečná, minimalizujúca akcia vždy existuje. Ak existuje taká akcia práve jedna, pravidlo je zjavne deterministické. Ak je takých akcií viac, možno ku každej zostrojiť deterministické pravidlo. Ďalej potom možno uvažovať len vybrané jedno z nich, pretože všetky vedú na ekvivalentné politiky s rovnakými výslednými stratami.
- Postupujeme spätnou rekurziou pre všetky ostatné časy. Takto optimalizujeme rozhodovacie pravidlá pre každý stav. Optimálne rozhodovacie pravidlo $\pi^*(a_t \mid s_{t-1})$ pripisuje jednotkovú pravdepodobnosť práve minimalizujúcej akcii pre daný čas $t \in \mathbb{T}$.
- V poslednom kroku pre čas $t = 1$ dostávame optimálne rozhodovacie pravidlá, na základe ktorých zostavíme optimálnu politiku $\pi^* \in \arg\min_{\pi \in \Pi} u^{\pi}(s_0)$.
- I keď optimálne rozhodovacie pravidlo v MRP nemusí byť určené jednoznačne, je vždy možné voliť ho ako deterministické. Preto je optimálna politika v MRP vždy deterministická.

1.3 Diskontovanie

Diskontovanie je proces, ktorý nám umožňuje vyjadriť rozdielnu dôležitosť blízkych a vzdialených krokov v rozhodovacom procese. Vzdialené dôsledky majú totiž spravidla pre agenta nižšiu váhu.

Rozdielnu dôležitosť pre blízke a vzdialené kroky v rozhodovacom procese kvantifikujeme pomocou diskontného faktora $\gamma_t \in [0, 1]$. Tento faktor je obvykle závislý na čase a s týmto časom sa zmenšuje. Táto časová závislosť je najčastejšie vyjadrená umocňovaním fixnej hodnoty parametru, tzn. $\gamma_t := \gamma^t$. Pomocou diskontného faktoru možno rozšíriť definíciu hodnotovej funkcie na tvar

$$u^\pi(s_{t-1}) := \mathrm{E}^\pi \left[\sum_{\substack{\tau \in \mathbb{T} \\ \tau \geq t}} \gamma_\tau \; l(s_\tau, a_\tau, s_{\tau-1}) \mid s_{t-1} = s \right] \text{ pre } \forall s \in \mathbb{S}. \tag{1.10}$$

Diskontovanie úzko súvisí so zastavovaním. Zastavenie je ekvivalentné skutočnosti, že od istého momentu majú pre nás ďalšie straty nulovú váhu. Nastavením hodnoty diskontného faktoru od istého momentu na 0 umožníme zastavenie rozhodovacieho procesu.

1.4 Plne pravdepodobnostný návrh

V tejto kapitole budeme uvažovať pre popis rozhodovacieho procesu rovnakú uzavretú slučku a predpoklady ako u MRP. Zhrňme už zadefinované pojmy.

Uvažujeme časy $t \in \mathbb{T}$, kde $\mathbb{T} = \{1, \ldots, H\}, H < +\infty$. Máme konečnú, neprázdnu množinu stavov $\mathbb{S}$ a konečnú, neprázdnu množinu akcií $\mathbb{A}$. Daný počiatočný stav s_0 podmieňuje všetky ostatné stavy. V každom čase agent volí akciu $a_t \in \mathbb{A}$ a posúva tak systém zo stavu $s_{t-1} \in \mathbb{S}$ do stavu $s_t \in \mathbb{S}$. Výber akcie $a_t \in \mathbb{A}$ je daný politikou agenta π. Politika π je postupnosť rozhodovacích pravidiel $\pi(a_t \mid s_{t-1})$.

Zmenou oproti MRP je v tejto kapitole spôsob, akým hodnotíme kvalitu politiky. Využijeme plne pravdepodobnostný návrh, na ktorý budeme odkazovať skratkou PPN. Jeho axiomatické zavedenie možno nájsť v [13]. Z tohto zavedenia taktiež vyplýva, že PPN je rozšírením MRP, ako možno ďalej nahliadnuť v [17].

V rámci MRP bolo kritériom pre optimálnu politiku minimalizovanie celkových očakávaných strát. To sa dá chápať aj ako snaha ovplyvniť hustotu pravdepodobnosti premenných uzavretej slučky tak, aby sa čo najviac priblížila požadovanému ideálu.

Definícia 1.4.1 (Chovanie uzavretej slučky)
Chovanie uzavretej slučky do horizontu $H \in \mathbb{N}$ je chápané ako súbor

$$b := (s_H, a_H, s_{H-1}, \ldots, s_1, a_1) \in \mathbb{B}. \tag{1.11}$$

Vďaka predpokladaným vlastnostiam množín $\mathbb{S}$ a $\mathbb{A}$ z definície 1.2.1 je aj množina $\mathbb{B}$ konečná.

Definícia 1.4.2 (Hustota pravdepodobnosti chovania uzavretej slučky)
Hustota pravdepodobnosti chovania uzavretej slučky $c^\pi(b)$ je $2H$-rozmerná hustota pravdepodobnosti, ktorá plne popisuje vlastnosti chovania $b \in \mathbb{B}$.
Túto hustotu možno faktorizovať ako [2]

$$c^\pi(b) = \prod_{t \in \mathbb{T}} p(s_t \mid a_t, s_{t-1}) \pi(a_t \mid s_{t-1}) := p(b)\pi(b). \tag{1.12}$$

Pri faktorizácií sme využili reťazové pravidlo (1.1). Tvar podmienok sa nám zjednodušil ako dôsledok Markovej vlastnosti predpokladanej u definícií 1.2.1 a 1.2.2. V náväznosti na MRP sme pre $\forall t \in \mathbb{T}$ označili podmienené hustoty pravdepodobnosti premennej s_t ako prechodové funkcie a podmienené hustoty pravdepodobnosti premennej a_t ako rozhodovacie pravidlá. Faktorizácia je korektná, keďže počiatočný stav s_0 je daný a je implicitnou súčasťou všetkých podmienok.

Prvý činiteľ faktorizácie $p(b) := \prod_{t \in \mathbb{T}} p(s_t \mid a_t, s_{t-1})$ je tvorený súčinom prechodových funkcií a popisuje známy a pevný model systému.

Druhý činiteľ faktorizácie $\pi(b) := \prod_{t \in \mathbb{T}} \pi(a_t \mid s_{t-1})$ je tvorený súčinom rozhodovacích pravidiel. Keďže je tento súčin určený politikou, budeme ho pre jednoduchosť značiť rovnakým písmenom ako samotnú politiku, tzn. $\pi(b)$.

Všetky MRP vedúce k rovnakej $c^\pi(b)$ sú ekvivalentné aj napriek prípadným odlišnostiam v stratových funkciách (viz [19]).

[2] PPN zväčša využíva iného značenia (viz [10], [13], [17], [18]). My sa však v náväznosti na kapitolu o MRP pridržíme už používaného.

Veta 1.4.1 (O marginálnej hustote pravdepodobnosti $c^\pi(s_{t-1})$)
Uvažujme konečný horizont $H \in \mathbb{T}$. Pre $\forall t \in \mathbb{T}$ platí, že marginálna hustota pravdepodobnosti $c^\pi(s_{t-1})$ je nezávislá na rozhodovacích pravidlách $\pi(a_\tau \mid s_{\tau-1})$ pre $\tau \geq t$.

Dôkaz. Označme kartézsky súčin $\mathbb{S}^H := \overbrace{\mathbb{S} \times \mathbb{S} \times \cdots \times \mathbb{S}}^{H\text{-krát}}$. Jeho prvkami sú súbory stavov $(s_H, s_{H-1}, \ldots, s_1)$. Obdobne zavádzame kartézsky súčin $\mathbb{A}^H$, ktorého prvkami sú súbory akcií. Marginálna hustota $c^\pi(s_{t-1})$ sa dá potom rozpísať ako

$$c^\pi(s_{t-1}) = \sum_{\substack{\mathbb{S}^{H-1} \\ \mathbb{A}^H}} \left(\prod_{\substack{\tau \in \mathbb{T} \\ \tau \geq t}} p(s_\tau \mid a_\tau, s_{\tau-1})\ \pi(a_\tau \mid s_{\tau-1}) \prod_{\substack{\tau \in \mathbb{T} \\ \tau < t}} p(s_\tau \mid a_\tau, s_{\tau-1})\ \pi(a_\tau \mid s_{\tau-1}) \right).$$

Podľa definície je v zátvorke súčin $2H$ podmienených hustôt pravdepodobnosti, ktoré popisujú chovanie uzavretej slučky. Pre výpočet marginály podľa definície 1.1.3 sčítame cez všetkých $2H-1$ premenných s výnimkou s_{t-1}.

Začnime sčítavanie cez $s_H \in \mathbb{S}$. Z celej zátvorky sa s_H nachádza len v argumente $p(s_H \mid a_H, s_{H-1})$, preto tieto pravdepodobnosti možno vysčítať na 1.

Pokračujme sčítavaním cez $a_H \in \mathbb{A}$. Akcia a_H podmieňuje $p(s_H \mid a_H, s_{H-1})$, avšak tento výraz už sa v súčine nenachádza, pretože sme ho v predchádzajúcom kroku vysčítali. Preto sa a_H nachádza výhradne v argumente $\pi(a_H \mid s_{H-1})$, pričom tieto pravdepodobnosti možno opäť vysčítať na 1. Takto pokračujeme vysčítavaním cez stavy a akcie s menšími a menšími indexami, až dokým nevysčítame cez $a_t \in \mathbb{A}$.

Keďže cez $s_{t-1} \in \mathbb{S}$ nesčítame, člen $p(s_{t-1} \mid a_{t-1}, s_{t-2})$ nevypadne. Z reťazového pravidla (1.1) dostávame, že $c^\pi(s_{t-1})$ je nezávislá na $\pi(a_\tau \mid s_{\tau-1})$ pre $\tau \geq t$ a závislosť na $\pi(a_\tau \mid s_{\tau-1})$ pre $\tau < t$ sa zachová. □

PPN zavádza hustotu pravdepodobnosti ideálneho chovania uzavretej slučky $c^i(b)$ ako

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

Hustota $c^i(b)$ popisuje preferencie agenta a nahradzuje vyjadrenie strát pomocou stratovej funkcie v MRP. O konštrukcii $c^i(b)$ detailnejšie pojednáva [20]. Obecne $c^i(b)$ priraďuje vysoké hodnoty požadovanému a nízke hodnoty nežiadúcemu chovaniu uzavretej slučky. Cieľom optimalizácie je priblíženie hustoty ideálneho a reálneho chovania agenta. Priblíženie sa hodnotí na základe Kullback-Leiblerovej divergencie (KLD), ktorej zavedenie možno nájsť v [21].

Definícia 1.4.3 (Kullback-Leiblerova divergencia)
Uvažujme konečnú množinu $\mathbb{B}$ a hustoty pravdepodobnosti $f(b), g(b)$, kde $g(b) > 0$. Blízkosť týchto hustôt možno určiť pomocou **Kullback-Leiblerovej divergencie (KLD)** definovanej ako

$$D(f||g) := \sum_{b \in \mathbb{B}} f(b) \ln \frac{f(b)}{g(b)}. \tag{1.14}$$

KLD nadobúda len nezáporné hodnoty. Maximálne priblíženie nastáva pre $f = g$, potom $D(f||g) = 0$. Dôkaz týchto vlastností možno nájsť v [21].

Hľadanie minima KLD v PPN je analógiou k hľadaniu minima celkových očakávaných strát v MRP. Optimálna politika je obdobne daná ako argument minimalizujúci KLD.

Definícia 1.4.4 (Optimálna politika v PPN)
Uvažujme chovanie uzavretej slučky $b \in \mathbb{B}$ popísané hustotou pravdepodobnosti $c^\pi(b)$ a ideálne chovanie popísané hustotou pravdepodobnosti $c^i(b)$. **Optimálna politika** π^* v PPN je daná ako

$$\pi^* \in \arg\min_{\pi \in \Pi} D\left(c^\pi || c^i\right). \tag{1.15}$$

Kvôli zjednodušeniu zápisu v nasledujúcom texte zavedieme analógiu k stratovej funkcii v MRP. Stratová funkcia v PPN vyjadruje straty v zmysle odchýlenia sa od ideálu. Označíme

$$l^\pi(s_t, a_t, s_{t-1}) := \ln \frac{p(s_t \mid a_t, s_{t-1})\pi(a_t \mid s_{t-1})}{p^i(s_t \mid a_t, s_{t-1})\pi^i(a_t \mid s_{t-1})}. \tag{1.16}$$

Indexovaním podľa π zdôrazňujeme, že v PPN politika vystupuje v definícií strát.

Čiastkové aj celkové straty sú v PPN definované analogicky ako v MRP vzťahmi (1.4) a (1.5). Opäť uvažujeme ich očakávané hodnoty. Do celkových očakávaných strát vstupuje politika v dôsledku definície $l^\pi(s_t, a_t, s_{t-1})$ nelineárne. To je kľúčový rozdiel oproti MRP.

Veta 1.4.2 (Súčtový tvar KLD)
KLD združených hustôt pravdepodobnosti $c^\pi(b)$ a $c^i(b)$ možno využitím marginálnej hustoty pravdepodobnosti $c^\pi(s_{t-1})$ prepísať do tvaru

$$D\left(c^\pi || c^i\right) = \sum_{t \in \mathbb{T}} \sum_{s_{t-1} \in \mathbb{S}} c^\pi(s_{t-1}) \sum_{\substack{s_t \in \mathbb{S} \\ a_t \in \mathbb{A}}} p(s_t \mid a_t, s_{t-1})\ \pi(a_t \mid s_{t-1})\ l^\pi(s_t, a_t, s_{t-1}). \tag{1.17}$$

Dôkaz. Dosadíme do KLD z definície hustôt a dostaneme

$$D\left(c^\pi || c^i\right) = \sum_{b \in \mathbb{B}} c^\pi(b) \ln \frac{c^\pi(b)}{c^i(b)} = \sum_{b \in \mathbb{B}} c^\pi(b) \ln \frac{\prod_{t \in \mathbb{T}} p(s_t \mid a_t, s_{t-1})\pi(a_t \mid s_{t-1})}{\prod_{t \in \mathbb{T}} p^i(s_t \mid a_t, s_{t-1})\pi^i(a_t \mid s_{t-1})}.$$

Využijeme prepis logaritmu súčinu na súčet logaritmov. Vďaka konečnosti množín $\mathbb{B}$ a $\mathbb{T}$ možno zameniť poradie súm a dostať

$$D\left(c^\pi || c^i\right) = \sum_{b \in \mathbb{B}} c^\pi(b) \sum_{t \in \mathbb{T}} \ln \frac{p(s_t \mid a_t, s_{t-1})\pi(a_t \mid s_{t-1})}{p^i(s_t \mid a_t, s_{t-1})\pi^i(a_t \mid s_{t-1})} = \sum_{t \in \mathbb{T}} \sum_{b \in \mathbb{B}} c^\pi(b) l^\pi(s_t, a_t, s_{t-1}).$$

Pre každé $t \in \mathbb{T}$ je príslušný člen sumy závislý len na premenných s_t, a_t, s_{t-1}. Z definície marginálnej hustoty pravdepodobnosti 1.1.3 máme

$$D\left(c^\pi || c^i\right) = \sum_{t \in \mathbb{T}} \sum_{\substack{s_t, s_{t-1} \in \mathbb{S} \\ a_t \in \mathbb{A}}} c^\pi(s_t, a_t, s_{t-1}) l^\pi(s_t, a_t, s_{t-1}).$$

Rozpíšeme $c^\pi(s_t, a_t, s_{t-1})$ reťazovým pravidlom (1.1) na

$$c^\pi(s_t, a_t, s_{t-1}) = c^\pi(s_t \mid a_t, s_{t-1}) c^\pi(a_t \mid s_{t-1}) c^\pi(s_{t-1}).$$

Opäť preznačíme podľa zaužívaného zvyku podmienené hustoty pravdepodobnosti premenných s_t a a_t. Vďaka konečnosti množín $\mathbb{S}$ a $\mathbb{A}$ možno zameniť poradie súm, čo spolu s nezávislosťou $c(s_{t-1})$ na s_t a a_t dáva celkovú rovnosť v tvare

$$D\left(c^\pi || c^i\right) = \sum_{t \in \mathbb{T}} \sum_{s_{t-1} \in \mathbb{S}} c^\pi(s_{t-1}) \sum_{\substack{s_t \in \mathbb{S} \\ a_t \in \mathbb{A}}} p(s_t \mid a_t, s_{t-1})\ \pi(a_t \mid s_{t-1}) l^\pi(s_t, a_t, s_{t-1}).$$

□

Vzťah pre optimálnu hodnotovú funkciu v PPN splýva s (1.7) z definície v MRP. Pripomenieme jej tvar

$$u^{\pi^*}(s_{t-1}) := \min_{\substack{\{\pi(a_\tau|s_{\tau-1})\} \\ \tau \geq t}} \mathrm{E}^{\pi} \left[\sum_{\substack{\tau \in \mathbb{T} \\ \tau \geq t}} l^{\pi}(s_\tau, a_\tau, s_{\tau-1}) \mid s_{t-1} = s \right] \text{ pre } \forall s \in \mathbb{S},$$

pričom dodefinujeme $u^{\pi^*}(s_H) := 0$.

Pre optimálnu politiku π^* v PPN, pri ktorej je dosahované minimum hodnotovej funkcie pre $\forall t \in \mathbb{T}$ platí

$$\min_{\pi \in \Pi} D\left(c^{\pi} || c^{i}\right) = D\left(c^{\pi^*} || c^{i}\right) = u^{\pi^*}(s_0).$$

Podrobnejší rozbor tohto tvrdenia možno nájsť v [22]. Optimálna politika π^* v PPN je obecne znáhodnená, nie deterministická.

Veta 1.4.3 (O optimálnej hodnotovej funkcii)
Uvažujme optimálnu hodnotovú funkciu u^{π^*}.
Potom pre $\forall t \in \mathbb{T}$ je $u^{\pi^*}(s_{t-1}) \geq 0$ a platí spätná funkčná rekurzia

$$u^{\pi^*}(s_{t-1}) = \min_{\{\pi(a_t|s_{t-1})\}} \sum_{\substack{s_t \in \mathbb{S} \\ a_t \in \mathbb{A}}} p(s_t \mid a_t, s_{t-1})\ \pi(a_t \mid s_{t-1}) \left(u^{\pi^*}(s_t) + l^{\pi}(s_t, a_t, s_{t-1}) \right). \tag{1.18}$$

Dôkaz. Skutočnosť, že pre $\forall t \in \mathbb{T}$ je $u^{\pi^*}(s_{t-1}) \geq 0$ plynie priamo z vlastností KLD. Rekurzívny vzťah dokážeme matematickou indukciou.

Uvažujme $t = H$. V tomto prípade sa v definícií hodnotovej funkcie minimalizácia aj suma cez čas redukujú na jediný člen pre $\tau = H$,

$$u^{\pi^*}(s_{H-1}) = \min_{\{\pi(a_H|s_{H-1})\}} \mathrm{E}^{\pi} \left[l^{\pi}(s_H, a_H, s_{H-1}) \mid s_{H-1} = s \right].$$

Keďže s_{H-1} je dané, pri vyjadrení očakávanej hodnoty sčítame len cez $s_H \in \mathbb{S}$ a $a_H \in \mathbb{A}$. Tvrdenie platí, pretože

$$u^{\pi^*}(s_{H-1}) = \min_{\{\pi(a_H|s_{H-1})\}} \sum_{\substack{s_H \in \mathbb{S} \\ a_H \in \mathbb{A}}} p(s_H \mid a_H, s_{H-1}) \pi(a_H \mid s_{H-1})\ \left(\underbrace{u^{\pi^*}(s_H)}_{0} + l^{\pi}(s_H, a_H, s_{H-1}) \right).$$

Uvažujme teraz $t-1 \in \mathbb{T}$ a pre dané s_{t-1} hodnotovú funkciu $u^{\pi^*}(s_{t-1})$. Vo výraze pre hodnotovú funkciu vyčleníme pre $\tau = t$ minimalizáciu aj príslušnú sumu. Dostávame rovnosť

$$u^{\pi^*}(s_{t-1}) = \min_{\{\pi(a_t|s_{t-1})\}} \min_{\substack{\{\pi(a_\tau|s_{\tau-1})\} \\ \tau > t}} \left(\sum_{\substack{s_t \in \mathbb{S} \\ a_t \in \mathbb{A}}} (*) + \sum_{\substack{\tau \in \mathbb{T} \\ \tau > t}} \sum_{\substack{s_\tau, s_{\tau-1} \in \mathbb{S} \\ a_\tau \in \mathbb{A}}} (**) \right),$$

kde výrazy v sumách majú tvar

$$\begin{aligned}
(*) &= p(s_t \mid a_t, s_{t-1})\ \pi(a_t \mid s_{t-1})\ l^{\pi}(s_t, a_t, s_{t-1}), \\
(**) &= p(s_\tau \mid a_\tau, s_{\tau-1}, s_{t-1})\ \pi(a_\tau \mid s_{\tau-1}, s_{t-1})\ p(s_{\tau-1} \mid s_{t-1})\ l^{\pi}(s_\tau, a_\tau, s_{\tau-1}) = \\
&= p(s_\tau \mid a_\tau, s_{\tau-1})\ \pi(a_\tau \mid s_{\tau-1})\ p(s_{\tau-1} \mid s_{t-1})\ l^{\pi}(s_\tau, a_\tau, s_{\tau-1}).
\end{aligned}$$

Výraz $(*)$ podľa vety 1.4.1 nie je ovplyvnený pravidlami $\pi(a_\tau \mid s_{\tau-1})$ pre $\tau > t$. Preto možno zameniť minimalizáciu so sumou,

$$u^{\pi^*}(s_{t-1}) = \min_{\{\pi(a_t|s_{t-1})\}} \left(\sum_{\substack{s_t \in \mathbb{S} \\ a_t \in \mathbb{A}}} (*) + \sum_{\substack{\tau \in \mathbb{T} \\ \tau > t}} \min_{\substack{\{\pi(a_\tau|s_{\tau-1})\} \\ \tau > t}} \sum_{\substack{s_\tau, s_{\tau-1} \in \mathbb{S} \\ a_\tau \in \mathbb{A}}} (**) \right).$$

Vysčítavaním od najväčšieho indexu $\tau = H$ spätne až po $\tau = t+1$ a užitím indukčného predpokladu dostávame

$$u^{\pi^*}(s_{t-1}) = \min_{\{\pi(a_t|s_{t-1})\}} \left(\sum_{\substack{s_t \in \mathbb{S} \\ a_t \in \mathbb{A}}} p(s_t \mid a_t, s_{t-1})\ \pi(a_t \mid s_{t-1}) l^\pi(s_t, a_t, s_{t-1}) + \sum_{s_t \in \mathbb{S}} p(s_t \mid s_{t-1}) u^{\pi^*}(s_t) \right).$$

Dosadením

$$p(s_t \mid s_{t-1}) = \sum_{a_t \in \mathbb{A}} p(s_t \mid a_t, s_{t-1}) \pi(a_t \mid s_{t-1})$$

získame konečnú podobu rekurzívneho vzťahu z tvrdenia vety. □

K optimálnej politike sa teda vieme vždy dostať vďaka spätnej funkčnej rekurzii hodnotovej funkcie ukázanej vo vete 1.4.3, pričom rekurziu začíname z $u^{\pi^*}(s_H) = 0$.

Predchádzajúce vety nám umožňujú nájsť optimálnu politiku postupom analogickým k dynamickému programovaniu.

Veta 1.4.4 (Hľadanie optimálnej politiky v PPN)
Uvažujme optimálnu hodnotovú funkciu $u^{\pi^*}(s_t) = -\ln w(s_t)$.
Potom pre $\forall t \in \mathbb{T}$ platí $w(s_t) \in (0, 1]$ a $w(s_t) \leq w(s_H) = 1$. Optimálne rozhodovacie pravidlá

$$\pi^*(a_t \mid s_{t-1}) = \pi^i(a_t \mid s_{t-1}) \frac{\exp\left[-\eta(a_t, s_{t-1})\right]}{w(s_{t-1})} \tag{1.19}$$

sa napočítajú pre $\forall t \in \mathbb{T}$ pomocou spätnej funkčnej rekurzie začínajúcej v $t = H$ podľa vzťahov

$$\eta(a_t, s_{t-1}) = \sum_{s_t \in \mathbb{S}} p(s_t \mid a_t, s_{t-1}) \ln \frac{p(s_t \mid a_t, s_{t-1})}{w(s_t) p^i(s_t \mid a_t, s_{t-1})}, \tag{1.20}$$

$$w(s_{t-1}) = \sum_{a_t \in \mathbb{A}} \pi^i(a_t \mid s_{t-1}) \exp\left[-\eta(a_t, s_{t-1})\right]. \tag{1.21}$$

Tieto rozhodovacie pravidlá tvoria optimálnu politiku π^* v PPN, ktorá splňuje

$$\min_{\pi \in \Pi} D(c^{\pi^*} || c^i) = -\ln w(s_0) = u^{\pi^*}(s_0).$$

Dôkaz. Obmedzenie hodnôt funkcie $w(s_t)$ plynie z jej definície pomocou optimálnej hodnotovej funkcie. Uvažujme teraz pravú stranu rekurzívneho vzťahu (1.18). Vnútornú zátvorku v tomto výraze možno rozpísať pomocou definícií $l^\pi(s_t, a_t, s_{t-1})$ z (1.16) a $w(s_t)$ z (1.21) na

$$-\ln w(s_t) + \ln \frac{p(s_t \mid a_t, s_{t-1})\pi(a_t \mid s_{t-1})}{p^i(s_t \mid a_t, s_{t-1})\pi^i(a_t \mid s_{t-1})} = \ln \frac{\pi(a_t \mid s_{t-1})}{\pi^i(a_t \mid s_{t-1})} + \ln \frac{p(s_t \mid a_t, s_{t-1})}{w(s_t)p^i(s_t \mid a_t, s_{t-1})}.$$

Dosadením tohto vyjadrenia pravá strana (1.18) prechádza na tvar

$$\min_{\{\pi(a_t|s_{t-1})\}} \sum_{\substack{s_t \in \mathbb{S} \\ a_t \in \mathbb{A}}} p(s_t \mid a_t, s_{t-1})\pi(a_t \mid s_{t-1}) \left(\ln \frac{\pi(a_t \mid s_{t-1})}{\pi^i(a_t \mid s_{t-1})} + \ln \frac{p(s_t \mid a_t, s_{t-1})}{w(s_t)p^i(s_t \mid a_t, s_{t-1})} \right).$$

Roznásobíme výrazy v zátvorkách. Prvý logaritmus nezávisí na premennej s_t, preto možno v tomto člene vysčítať pravdepodobnosti $p(s_t \mid a_t, s_{t-1})$ na 1 a dostať

$$\min_{\{\pi(a_t|s_{t-1})\}} \sum_{a_t \in \mathbb{A}} \pi(a_t \mid s_{t-1}) \left(\ln \frac{\pi(a_t \mid s_{t-1})}{\pi^i(a_t \mid s_{t-1})} + \sum_{s_t \in \mathbb{S}} p(s_t \mid a_t, s_{t-1}) \ln \frac{p(s_t \mid a_t, s_{t-1})}{w(s_t)p^i(s_t \mid a_t, s_{t-1})} \right).$$

Využitím (1.20) upravíme celú rovnosť (1.18) na

$$-\ln w(s_{t-1}) = \min_{\{\pi(a_t|s_{t-1})\}} \sum_{a_t \in \mathbb{A}} \pi(a_t \mid s_{t-1}) \ln \frac{\pi(a_t \mid s_{t-1})}{\pi^i(a_t \mid s_{t-1}) \exp\left[-\eta(a_t, s_{t-1})\right]}.$$

Prevedením $\ln w(s_{t-1})$ na pravú stranu normalizujeme menovateľ logaritmu,

$$0 = \min_{\{\pi(a_t|s_{t-1})\}} \sum_{a_t \in \mathbb{A}} \pi(a_t \mid s_{t-1}) \ln \frac{\pi(a_t \mid s_{t-1})}{\pi^i(a_t \mid s_{t-1}) \frac{\exp\left[-\eta(a_t, s_{t-1})\right]}{w(s_{t-1})}}.$$

Dostávame KLD, ktorá je najmenšia (nulová) práve pre optimálne rozhodovacie pravidlá

$$\pi^*(a_t \mid s_{t-1}) = \pi^i(a_t \mid s_{t-1}) \frac{\exp\left[-\eta(a_t, s_{t-1})\right]}{w(s_{t-1})}.$$

□

Kapitola 2

Problematika zastavovania v PPN

2.1 Úvodné predpoklady

Táto kapitola predstavuje jadro práce, v ktorej rozširujeme PPN o zastavovanie a diskutujeme popis zastavovacích pravidiel. V závere kapitoly užijeme vetu 1.4.4 na takto rozšírený popis, načrtneme riešenie optimálnej politiky v rámci PPN so zastavovaním a poukážeme na otvorené problémy. Nasledujúce predpoklady sú kľúčové:

- Ak sa raz proces zastaví, toto zastavenie je trvalé.
- Zastavenie má dvojitú povahu.
  Za prvé, zastavenie má podobu akcie. Táto akcia ovplyvňuje celé budúce chovanie uzavretej slučky. Za druhé, zastavenie je súčasťou stavu v popise modelu systému.
- Daný počiatočný stav s_0 doplníme predpokladom, že v čase $t = 0$ nezastavujeme.

Definícia 2.1.1 (Akcia predĺženia)
Akcia predĺženia je definovaná pre $\forall t \in \mathbb{T}$ ako

$$\tilde{a}_t := \begin{cases} 1 & \text{pre pokračovanie v generovaní } a_t, \\ 0 & \text{pre zastavenie generovania } a_\tau \text{ pre } \forall \tau \geq t. \end{cases}$$

Množina akcií predĺženia je dvojprvková, $\tilde{\mathbb{A}} = \{0, 1\}$.

Kvôli jednoznačnému rozlíšeniu medzi dvoma typmi uvažovaných akcií budeme v ďalšom texte akciu $a_t \in \mathbb{A}$ nazývať normálna akcia.

Zavedením akcie predĺženia rozširujeme normálnu akciu a_t na dvojicu $(a_t, \tilde{a}_t)$. Dvojitú povahu zastavenia rozlíšime v popise modelu systému tak, že stav s_t rozšírime na dvojicu (s_t, z_t). Zložka stavu $z_t \in \{0, 1\}$ popisuje, či v danom čase je proces zastavený ($z_t = 0$), alebo nie ($z_t = 1$). Podľa úvodných predpokladov dodefinujeme $z_0 := 1$.

Pre každý čas $t \in \mathbb{T}$ sledujeme vývoj $a_t, \tilde{a}_t, s_{t-1}, z_{t-1} \to s_t, z_t$. Chovanie uzavretej slučky sa uvažovaním zastavenia rozširuje na usporiadaný súbor $4H$ členov

$$b := (s_H, z_H, a_H, \tilde{a}_H, \ldots, s_1, z_1, a_1, \tilde{a}_1) \in \mathbb{B}. \tag{2.1}$$

Hustotu pravdepodobnosti pre chovanie uzavretej slučky možno opäť faktorizovať ako

$$c(b) = \prod_{t \in \mathbb{T}} p(s_t, z_t \mid a_t, \tilde{a}_t, s_{t-1}, z_{t-1}) \pi(a_t, \tilde{a}_t \mid s_{t-1}, z_{t-1}) := p(b)\pi(b). \tag{2.2}$$

2.2 Rozšírenie rozhodovacích pravidiel

Akcia predĺženia rozširuje popis rozhodovacích pravidiel o zastavovacie pravidlá. Užitím reťazového pravidla (1.1) dostávame [1]

$$\pi(a_t, \tilde{a}_t \mid s_{t-1}, z_{t-1}) = \pi(a_t \mid \tilde{a}_t, s_{t-1}, z_{t-1})\pi(\tilde{a}_t \mid s_{t-1}, z_{t-1}). \tag{2.3}$$

Prvý činiteľ v (2.3) definujeme vzťahom

$$\pi(a_t \mid \tilde{a}_t, s_{t-1}, z_{t-1}) := \Big(\pi(a_t \mid s_{t-1})\Big)^{\tilde{a}_t}\Big(\tilde{\pi}(a_t \mid s_{t-1})\Big)^{1-\tilde{a}_t}. \tag{2.4}$$

Rozhodovacie pravidlo $\pi(a_t \mid s_{t-1})$ je totiž ďalej optimalizované práve vtedy, keď $\tilde{a}_t = 1$. Ako vyplynie z nasledujúceho textu, rozhodovacie pravidlo $\tilde{\pi}(a_t \mid s_{t-1})$ je nadbytočné, pretože jeho vplyv sa po zastavení procesu reálne neprejaví. Možno ho voliť podľa ľubovoľnej politiky.

Diskutujme ďalej druhý činiteľ v (2.3). Pre zastavenie v stave s_{t-1} je akcia predĺženia $\tilde{a}_{t-1} = 0$. Normálna akcia a_{t-1} sa negeneruje a $z_{t-1} = 0$. Pripomíname, že uvažujeme predpoklad, kedy je zastavenie trvalé. Preto ak $\tilde{a}_{t-1} = 0$, je $\tilde{a}_\tau = 0$ pre $\forall \tau \geq t$ s pravdepodobnosťou 1. Pre pokračovanie v stave s_{t-1} je akcia predĺženia $\tilde{a}_{t-1} = 1$. Do tohto stavu prechádzame normálnou akciou a_{t-1}, pričom $z_{t-1} = 1$. V tomto stave posúdime, či chceme naďalej pokračovať v procese. Pre $z_{t-1} = 1$ zavedieme pravdepodobnosť predĺženia $q(s_{t-1})$, ktorá závisí len na danom stave, v ktorom ďalšie pokračovanie v procese zvažujeme. Naše rozhodnutie sa premieta do akcie $\tilde{a}_t$. Druhý činiteľ v (2.3) predstavujúci zastavovacie pravidlo preto uvažujeme výhradne v tvare

$$\pi(\tilde{a}_t \mid s_{t-1}, z_{t-1}) := \begin{cases} 1 & \text{pre } \tilde{a}_t = 0 \text{ ak } z_{t-1} = 0, \\ 0 & \text{pre } \tilde{a}_t = 1 \text{ ak } z_{t-1} = 0, \\ 1 - q(s_{t-1}) & \text{pre } \tilde{a}_t = 0 \text{ ak } z_{t-1} = 1, \\ q(s_{t-1}) & \text{pre } \tilde{a}_t = 1 \text{ ak } z_{t-1} = 1. \end{cases} \tag{2.5}$$

2.3 Rozšírenie prechodových funkcií

Užime reťazové pravidlo (1.1) na hustotu pravdepodobnosti $p(s_t, z_t \mid a_t, \tilde{a}_t, s_{t-1}, z_{t-1})$ z (2.2). Dostávame výraz $p(z_t \mid a_t, \tilde{a}_t, s_{t-1}, z_{t-1})$, tzn. prechodovú funkciu pre zastavovaciu zložku stavu. Tú definujeme pomocou Kroneckerovho delta

$$p(z_t \mid a_t, \tilde{a}_t, s_{t-1}, z_{t-1}) := \delta_{\tilde{a}_t, z_t} = \begin{cases} 1 & \text{ak } \tilde{a}_t = z_t, \\ 0 & \text{inak.} \end{cases}$$

Predpokladáme, že akcie predĺženia nemenia pravdepodobnosť prechodu medzi jednotlivými stavmi. Nový stav s_t závisí len na predchádzajúcom stave s_{t-1} a zvolenej normálnej akcii a_t. Preto možno zjednodušiť tvar $p(s_t \mid z_t, a_t, \tilde{a}_t, s_{t-1}, z_{t-1})$ na $p(s_t \mid a_t, s_{t-1})$ a celkovo dostávame

$$p(s_t, z_t \mid a_t, \tilde{a}_t, s_{t-1}, z_{t-1}) = p(s_t \mid z_t, a_t, \tilde{a}_t, s_{t-1}, z_{t-1})p(z_t \mid a_t, \tilde{a}_t, s_{t-1}, z_{t-1}) = p(s_t \mid a_t, s_{t-1})\delta_{\tilde{a}_t z_t}.$$

[1] V rámci rozhodovacích pravidiel podmieňujeme normálne akcie a_t, v rámci zastavovacích pravidiel podmieňujeme akcie predĺženia $\tilde{a}_t$.

2.4 Návrh nových ideálov

Zostáva nám diskutovať popis ideálnej hustoty pravdepodobnosti chovania uzavretej slučky so zastavením.

Do momentu zastavenia musia byť ideálny model systému rovnako ako ideálna politika identické s ideálmi z popisu bez zastavení, ktoré sme v (1.13) označili ako $p^i(s_t \mid a_t, s_{t-1})$ a $\pi^i(a_t \mid s_{t-1})$.

Od momentu zastavenia by sa nemala meniť hodnota minima KLD. Formálne je najjednoduchšie využiť *"prenechanie voľby osudu"* (viz [17]) a stotožniť ideály s $p(s_t \mid a_t, s_{t-1})$ resp. $\tilde{\pi}(a_t \mid s_{t-1})$.

Ideálne zastavovacie pravidlá $\pi^i(\tilde{a}_t \mid s_{t-1}, z_{t-1})$ sú určené v zmysle (2.5) využitím ideálnej pravdepodobnosti predĺženia $q^i(s_{t-1})$. Z týchto úvah celkovo dostávame

$$p^i(s_t, z_t \mid a_t, \tilde{a}_t, s_{t-1}, z_{t-1}) := \Big(p^i(s_t \mid a_t, s_{t-1})\Big)^{\tilde{a}_t} \Big(p(s_t \mid a_t, s_{t-1})\Big)^{1-\tilde{a}_t} \delta_{\tilde{a}_t z_t}, \tag{2.6}$$

$$\pi^i(a_t, \tilde{a}_t \mid s_{t-1}, z_{t-1}) := \Big(\pi^i(a_t \mid s_{t-1})\Big)^{\tilde{a}_t} \Big(\tilde{\pi}(a_t \mid s_{t-1})\Big)^{1-\tilde{a}_t} \pi^i(\tilde{a}_t \mid s_{t-1}, z_{t-1}), \tag{2.7}$$

$$\pi^i(\tilde{a}_t \mid s_{t-1}, z_{t-1}) := \begin{cases} 1 & \text{pre } \tilde{a}_t = 0 \text{ ak } z_{t-1} = 0, \\ 0 & \text{pre } \tilde{a}_t = 1 \text{ ak } z_{t-1} = 0, \\ 1 - q^i(s_{t-1}) & \text{pre } \tilde{a}_t = 0 \text{ ak } z_{t-1} = 1, \\ q^i(s_{t-1}) & \text{pre } \tilde{a}_t = 1 \text{ ak } z_{t-1} = 1. \end{cases} \tag{2.8}$$

2.5 Hľadanie optimálnej politiky v PPN so zastavovaním

Vo vete 1.4.4 sme odvodili rekurzívne vzťahy

$$\eta(a_t, s_{t-1}) = \sum_{s_t \in \mathbb{S}} p(s_t \mid a_t, s_{t-1}) \ln \frac{p(s_t \mid a_t, s_{t-1})}{w(s_t) p^i(s_t \mid a_t, s_{t-1})},$$

$$w(s_{t-1}) = \sum_{a_t \in \mathbb{A}} \pi^i(a_t \mid s_{t-1}) \exp\big[-\eta(a_t, s_{t-1})\big].$$

Tieto rekurzívne vzťahy určujú optimálnu politiku danú optimálnymi rozhodovacími pravidlami

$$\pi^*(a_t \mid s_{t-1}) = \pi^i(a_t \mid s_{t-1}) \frac{\exp\big[-\eta(a_t, s_{t-1})\big]}{w(s_{t-1})}.$$

Pre rozšírené stavy a akcie potrebujeme $\eta(a_t, \tilde{a}_t, s_{t-1}, z_{t-1})$ a $w(s_{t-1}, z_{t-1})$. Ako možno nahliadnuť z (2.3), rozšírením dostaneme optimálne rozhodovacie pravidlá $\pi^*(a_t, \tilde{a}_t \mid s_{t-1}, z_{t-1})$ v tvare

$$\Big(\pi^i(a_t \mid s_{t-1})\Big)^{\tilde{a}_t} \Big(\tilde{\pi}(a_t \mid s_{t-1})\Big)^{1-\tilde{a}_t} \pi^i(\tilde{a}_t \mid s_{t-1}, z_{t-1}) \frac{\exp\big[-\eta(a_t, \tilde{a}_t, s_{t-1}, z_{t-1})\big]}{w(s_{t-1}, z_{t-1})}. \tag{2.9}$$

Činiteľ $\pi^i(\tilde{a}_t \mid s_{t-1}, z_{t-1})$ z (2.8) zaručuje, že pre $z_{t-1} = 0$ je nulová pravdepodobnosť, aby $\tilde{a}_t = 1$. Hodnota minimalizácie KLD sa nemení. Uvažujme preto ďalej $z_{t-1} = 1$. Budeme rozlišovať dva prípady, zvlášť pre zastavenie $\tilde{a}_t = 0$ a predĺženie $\tilde{a}_t = 1$.

2.6 Prípad zastavenia, $\tilde{a}_t = 0$

Pre prípad zastavenia dostávame

$$\begin{aligned}\eta(a_t, \tilde{a}_t = 0, s_{t-1}, z_{t-1} = 1) &= -\sum_{s_t \in \mathbb{S}} p(s_t \mid a_t, s_{t-1}) \ln w(s_t, z_t = 0) \\ &= \sum_{s_t \in \mathbb{S}} p(s_t \mid a_t, s_{t-1}) u^{\pi^*}(s_t, z_t = 0) \\ &= 0.\end{aligned}$$

V poslednom kroku využívame, že optimálna hodnotová funkcia $u^{\pi^*}(s_t, z_t = 0) = -\ln w(s_t, z_t = 0)$ vyjadruje KLD hustôt po zastavení. Toto zastavenie je trvalé. V tomto prípade ale príslušný faktor $c^i(b)$ splýva s $c^{\pi^*}(b)$, čo odpovedá nulovej hodnote KLD.

Z nezápornosti KLD a z faktu, že dosiahnutie horizontu vynucuje zastavenie dostávame $0 \leq w(s_t, z_t) \leq w(s_H, z_H = 0) = 1$ pre $\forall t \in \mathbb{T}$. Obecnejšie, $w(s_t, z_t) = 1$ pre $\forall t \in \mathbb{T}$ od momentu zastavenia.

2.7 Prípad predĺženia, $\tilde{a}_t = 1$

Pre prípad predĺženia dostávame

$$\eta(a_t, \tilde{a}_t = 1, s_{t-1}, z_{t-1} = 1) = \sum_{s_t \in \mathbb{S}} p(s_t \mid a_t, s_{t-1}) \ln \frac{p(s_t \mid a_t, s_{t-1})}{w(s_t, z_t = 1) p^i(s_t \mid a_t, s_{t-1})}.$$

Pri hľadaní optimálnej politiky potrebujeme napočítať rekurziu

$$w(s_{t-1}, z_{t-1} = 1) = \sum_{\tilde{a}_t \in \tilde{\mathbb{A}}} \sum_{a_t \in \mathbb{A}} \pi^i(a_t, \tilde{a}_t \mid s_{t-1}, z_{t-1} = 1) \exp\big[-\eta(a_t, \tilde{a}_t, s_{t-1}, z_{t-1} = 1)\big] = (*).$$

Výraz $(*)$ možno rozpísať do tvaru

$$\begin{aligned}(*) &= \sum_{a_t \in \mathbb{A}} \Big(1 - q^i(s_{t-1})\Big) \tilde{\pi}(a_t \mid s_{t-1}) \exp\big[-\eta(a_t, \tilde{a}_t = 0, s_{t-1}, z_{t-1} = 1)\big] + \\ &+ \sum_{a_t \in \mathbb{A}} q^i(s_{t-1})\ \pi^i(a_t \mid s_{t-1})\ \exp\big[-\eta(a_t, \tilde{a}_t = 1, s_{t-1}, z_{t-1} = 1)\big] = \\ &= \Big(1 - q^i(s_{t-1})\Big) \sum_{a_t \in \mathbb{A}} \tilde{\pi}(a_t \mid s_{t-1}) + \\ &+ q^i(s_{t-1}) \sum_{a_t \in \mathbb{A}} \pi^i(a_t \mid s_{t-1})\ \exp\big[-\eta(a_t, \tilde{a}_t = 1, s_{t-1}, z_{t-1} = 1)\big] = \\ &= 1 - q^i(s_{t-1}) + q^i(s_{t-1}) \sum_{a_t \in \mathbb{A}} \pi^i(a_t \mid s_{t-1})\ \exp\big[-\eta(a_t, \tilde{a}_t = 1, s_{t-1}, z_{t-1} = 1)\big].\end{aligned}$$

Využili sme, že podľa vzťahov (2.7) a (2.8)

$$\begin{aligned}\pi^i(a_t, \tilde{a}_t = 0 \mid s_{t-1}, z_{t-1} = 1) &= \Big(1 - q^i(s_{t-1})\Big) \tilde{\pi}(a_t \mid s_{t-1}), \\ \pi^i(a_t, \tilde{a}_t = 1 \mid s_{t-1}, z_{t-1} = 1) &= q^i(s_{t-1}) \pi^i(a_t \mid s_{t-1}).\end{aligned}$$

2.8 Riešenie optimálnej politiky v PPN so zastavovaním

Aplikovaním vety 1.4.4 na rozšírený popis PPN so zastavovaním dostávame tvrdenie, ktoré má nasledujúcu štruktúru.

Lemma 2.8.1 (Hľadanie optimálnej politiky v PPN so zastavovaním)
Uvažujme optimálnu hodnotovú funkciu $u^{\pi^*}(s_t, z_t) = -\ln w(s_t, z_t)$.
Potom pre $\forall t \in \mathbb{T}$ platí $w(s_t, z_t) \in (0, 1]$ a $w(s_t, z_t) \leq w(s_H, z_H = 0) = 1$. Uvažujme navyše čas zastavenia $t_0 \in \mathbb{T}$ a ideály prechodových funkcií, rozhodovacích pravidiel a zastavovacích pravidiel z (2.6), (2.7) a (2.8).

Optimálne rozhodovacie pravidlá pre $\forall t < t_0$ v tvare

$$\pi^*(a_t, \tilde{a}_t = 1 \mid s_{t-1}, z_{t-1} = 1) = \pi^i(a_t \mid s_{t-1}) q^i(s_{t-1}) \frac{\exp\left[-\eta(a_t, \tilde{a}_t = 1, s_{t-1}, z_{t-1} = 1)\right]}{w(s_{t-1}, z_{t-1} = 1)} \quad (2.10)$$

sa napočítajú pomocou spätnej funkčnej rekurzie začínajúcej v t_0 podľa vzťahov

$$\eta(a_t, \tilde{a}_t = 1, s_{t-1}, z_{t-1} = 1) = \sum_{s_t \in \mathbb{S}} p(s_t \mid a_t, s_{t-1}) \ln \frac{p(s_t \mid a_t, s_{t-1})}{w(s_t, z_t = 1) p^i(s_t \mid a_t, s_{t-1})}, \quad (2.11)$$

$$w(s_{t-1}, z_{t-1} = 1) = 1 - q^i(s_{t-1}) + q^i(s_{t-1}) \sum_{a_t \in \mathbb{A}} \pi^i(a_t \mid s_{t-1}) \ \exp\left[-\eta(a_t, \tilde{a}_t = 1, s_{t-1}, z_{t-1} = 1)\right]. \quad (2.12)$$

Pre $t = t_0$ je $w(s_{t_0}, z_{t_0} = 0) = 1$ a optimálne rozhodovacie pravidlo sa mení na

$$\pi^*(a_t, \tilde{a}_t = 0 \mid s_{t-1}, z_{t-1} = 1) = \pi(a_t \mid s_{t-1}) \frac{1 - q^i(s_{t-1})}{w(s_{t-1}, z_{t-1} = 1)}. \quad (2.13)$$

Pre $t > t_0$ je $w(s_t, z_t = 0) = 1$, pričom optimálne rozhodovacie pravidlo je volené ľubovoľne.
Uvedené rozhodovacie pravidlá tvoria optimálnu politiku π^* v PPN so zastavovaním, ktorá splňuje

$$\min_{\Pi} D(c^{\pi^*} || c^i) = -\ln w(s_0, z_0 = 1) = u^{\pi^*}(s_0, z_0 = 1). \quad (2.14)$$

Kľúčové rozšírenie vety 1.4.4 spočíva vo vzťahoch (2.11) a (2.12). Pre voľbu $q^i(s_{t-1}) = 1$ dostávame popis bez zastavenia.

Problematickým bodom tvrdenia 2.8.1 je skutočnosť, že potrebujeme explicitne poznať čas zastavenia t_0. Avšak tento čas by mal z rozšírenia teórie sám vyplynúť, tzn. na základe teórie by malo byť možné porovnať rôzne časy zastavenia a vybrať ten najvhodnejší (minimalizujúci stratu). Uvedená veta je preto pre riešenie optimalizačných úloh pomocou PPN so zastavovaním prakticky nepoužiteľná.

Kľúčový nedostatok diskutovaného rozšírenia PPN o zastavovanie je pravdepodobne v návrhu ideálu (2.6). Po aplikovaní reťazového pravidla (1.1) na tento ideál dostaneme člen $p^i(z_t \mid a_t, \tilde{a}_t, s_{t-1}, z_{t-1})$, ktorý by zrejme nemal byť položený Kroneckerovmu delta, ako sme to urobili v prípade $p(z_t \mid a_t, \tilde{a}_t, s_{t-1}, z_{t-1})$. Člen ideálu by mal mať sofistikovanejší tvar. Ideál so zastavovacou zložkou by totiž mal vyjadrovať dosadením do KLD dva typy strát. Za prvé, stratu kumulovanú v dôsledku toho, že sme proces ešte nezastavili a robíme ďalšie pozorovania. Za druhé, stratu v dôsledku konečného zastavenia, tzn. preklopenia $z_{t-1} \to z_t \neq z_{t-1}$. Tieto straty budú v každom momente ovplyvnené povahou konkrétneho problému, budú teda obecne závislé na $a_t \in \mathbb{A}$ a $s_t \in \mathbb{S}$. Konečná podoba tohto ideálu by mala byť predmetom ďalšieho skúmania.

Záver

Táto bakalárska práca sa zaoberá problematikou zastavovania v teórií rozhodovacích procesov. Najprv sú zavedené základné pojmy z teórie pravdepodobnosti, ktoré sú potrebné pre diskrétne Markove rozhodovacie procesy (MRP) a ich rozšírenie v podobe plne pravdepodobnostného návrhu (PPN). Rozšírenie v podobe PPN umožňuje riešiť dynamické úlohy.

Prvým prínosom tejto práce je zozbieranie pomocných tvrdení potrebných pre nájdenie optimálnej politiky v PPN, zjednotenie značenia v nich a ich dokázanie. Takto je v jasnej nadväznosti na MRP od základu odvodená veta, ktorá je analógiou dynamického programovania pre PPN. Táto veta je hlavným nástrojom tejto práce.

Ďalším prínosom je zavedenie PPN so zastavovaním novou, systematickejšou cestou. Kvôli dvojitej povahe zastavenia je zavedená akcia predĺženia, aj zastavovacia zložka stavu. Ďalej je diskutovaný tvar zastavovacích pravidiel a tvar členov hustoty pravdepodobnosti, ktorá popisuje ideálne chovanie uzavretej slučky so zastavovaním. Tvary týchto členov sa prvýkrát objavujú v tejto práci, čím práca významne prispieva k rozšíreniu PPN. Veta, ktorá je analógiou dynamického programovania pre PPN, je následne užitá na takto rozšírený popis. Aplikovanie vety odhaľuje problém, kedy takto rozšírený popis PPN so zastavovaním stále nedáva uspokojivú odpoveď pri hľadaní optimálneho času zastavenia. Predpokladá sa, že problém spočíva v nedostatočne kvalitnom návrhu ideálu pre model systému.

Súčasťou práce bola tvorba programov pre osvojenie si riešenia jednoduchých optimalizačných úloh pomocou MRP, bayesovského odhadovania a dynamického programovania v PPN. Bol rozpracovaný návrh experimentu, ktorý sa zaoberá odhadom neznámeho parametru využitím bayesovského učenia a PPN so zastavovaním. Vzhľadom na náročnosť teoretickej časti sa nepodarilo dostatočne zapracovať problematiku zastavovania do programov simulujúcich PPN. Z týchto dôvodov experimentálna časť nie je súčasťou tejto práce.

V nadväzujúcej práci by bolo vhodné opätovné diskutovanie ideálov v PPN so zastavovaním. Po následnej realizácií navrhnutého experimentu by bolo možné porovnať výsledky s klasickým prístupom pri odhadovaní parametrov. Je tiež možné navrhnúť iné experimenty pre ilustrovanie rôznych foriem zastavovania. Tie by napr. mohli zohľadňovať zastavenie len časti rozhodovacieho procesu (napr. zastavenie odhadu istého parametru v rámci učenia sa modelu systému), ale pokračovať v optimalizácií normálnych akcií. Otvoreným problémom v PPN je súvis medzi zastavovaním a diskontovaním, ktorý je načrtnutý v kapitole 1.3.

Register

Bibliografia

[1] M. Puterman. *Markov decision processes*. John Wiley a Sons, 1994, s. 325 –336.

[2] P.C. Fishburn. „Nontransitive Preferences in Decision Theory“. In: *Journal of Risk and Uncertainty* 4 (1991), s. 113–134.

[3] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 1985.

[4] A. Wald. *Sequential Analysis*. Dover Publications, 2013, s. 1 –54.

[5] M. Buckmann a M. Fifić. „Stopping Rule Selection (SRS) Theory Applied to Deferred Decision Making“. In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 35 (2013).

[6] T. S. Ferguson. „Who solved the secretary problem?“ In: *Statistical Science* 4.3 (1989), s. 282 –296.

[7] T. J. Lorenzen. „Optimal stopping with sampling cost: The secretary problem“. In: *The Annals of Probability* 9 (1981), s. 167 –172.

[8] A. Rapoport a D. A. Seale. „Optimal Stopping Behavior with RelativeRanks: The Secretary Problem with Unknown Population Size“. In: *Behavioral Decision Making* 13 (2000), s. 391 –411.

[9] L. J. Savage. *The Foundations of Statistics*. New York: Wiley, 1954, s. 239 –304.

[10] T.V. Guy, M. Kárný a D.H. Wolpert. *Decision Making: Uncertainty, Imperfection, Deliberation and Scalability*. Springer, 2014, s. 57–92.

[11] T.V. Guy, M. Kárný a D.H. Wolpert. *Decision Making and Imperfection*. Springer-Verlag (Studies in Computational Intelligence), 2013. DOI: 10.1007/978-3-642-36406-8.

[12] T.V. Guy, M. Kárný a D.H. Wolpert. *Decision Making with Imperfect Decision Makers*. Springer-Verlag (Intelligent Systems Reference Library), 2012. DOI: 10.1007/978-3-642-24647-0.

[13] M. Kárný. „Axiomatisation of fully probabilistic design revisited“. In: *Systems & Control Letters* 141 (2020).

[14] J. Jacod a P. Protter. *Probability Essentials*. Berlin Heidelberg: Springer, 2004.

[15] A. Rényi. *Probability Theory*. Dover Publications, 1970.

[16] R. Bellman. *Dynamic Programming*. N.Y.: Princeton University Press, 1957.

[17] J. Böhm et al. *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. Springer, 2006.

[18] T.V. Guy a M. Kárný. „Fully probabilistic control design“. In: *Systems & Control Letters* 55.4 (2006), s. 259–265.

[19] M. Ullrich. „Optimum control of some stochastic systems“. In: *Interní správy Akademie věd České republiky, Ústav teorie informace a automatizace, v.v.i.* (1964).

[20] T. Sivakova. „Algoritmický výběr dosažitelných preferencí“. In: *Interní správy Akademie věd České republiky, Ústav teorie informace a automatizace, v.v.i.* (2020).

[21] S. Kullback a R. Leibler. „On information and sufficiency“. In: *Ann Math Stat* 22 (1951), s. 79 –87.

[22] M. Kárný, J. Šindelář a I. Vajda. „Stochastic control optimal in the Kullback sense“. In: *Kybernetika* 44.1 (2008), s. 53–60.