České vysoké učení technické v Praze
Fakulta jaderná a fyzikálně inženýrská

Katedra matematiky
Obor: Matematické inženýrství

Bayesovské odhadování pro adaptivní dynamické rozhodování

Bayesian estimation for adaptive dynamic decision making

BAKALÁŘSKÁ PRÁCE

Vypracoval: Jurij Ružejnikov
Vedoucí práce: Dipl.-Eng. Tatiana V. Guy, PhD.
Rok: 2021

Katedra: matematiky

Akademický rok: 2020/2021

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Student:	Jurij Ružejnikov
Studijní program:	Aplikace přírodních věd
Studijní obor:	Matematické inženýrství
Studijní zaměření:	Aplikované matematicko-stochastické metody
Název práce (česky):	Bayesovské odhadování pro adaptivní dynamické rozhodování
Název práce (anglicky):	Bayesian estimation for adaptive dynamic decision making

Pokyny pro vypracování:

1) Najděte motivující aplikaci (příklad), kde je potřeba adaptivní dynamické rozhodování [4].
2) Naformulujte rozhodovací úlohu pro vybraný příklad. Seznamte se s dynamickým programováním [1] pro markovské rozhodovací procesy [2].
3) Seznamte se s bayesovským odhadováním v uzavřené smyčce [3] pro diskretní případ jak v čase, tak i v hodnotách.
4) Vytvořte algoritmus realizující odhadování modelu dle bodu 3. a implementujte ho v prostředí Matlab nebo Python.
5) Experimentálně vyhodnoťte kvalitu Vaší předpovědi na simulovaných nebo reálných datech.

Doporučená literatura:

1) R. E. Bellman, Dynamic Programming. Princeton University Press, Princeton, 1957. (vybrané části)

2) V. Peterka, Bayesian System Identification. In 'P. Eykhoff: Trends and Progress in System Identification', Pergamon Press, Oxford, 1981, 239--304. (vybrané části)

3) M. Puterman, Markov decision processes. John Wiley \& Sons, 1994. (vybrané části)

4) W. Powell, From Reinforcement Learning to Optimal Control: A unified framework for sequential decisions. ArXiv 1912.03513, 2019.

Jméno a pracoviště vedoucího bakalářské práce:

Dipl.-Eng. Tatiana V. Guy, PhD.
ÚTIA AVČR, v.v.i., Pod vodárenskou věží 4, 182 00 Praha 8

Jméno a pracoviště konzultanta:

Ing. Marko Ruman
ÚTIA AVČR, v.v.i., Pod vodárenskou věží 4, 182 00 Praha 8

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

Prohlášení

Prohlašuji, že jsem svou bakalářskou práci vypracoval samostatně a použil jsem pouze podklady (literaturu, projekty, SW atd.) uvedené v přiloženém seznamu.

V Praze dne .....

.....
Jurij Ružejnikov

Poděkování

Děkuji Dipl.-Eng. Tatiana V. Guy, PhD. za neocenitelné rady a pomoc při tvorbě bakalářské práce. Taktéž děkuji za podněty Ing. Marko Rumanovi a Ing. Františkovi Hůlovi.
Práce byla částečně podpořena projektem MŠMT LTC 18075.

Jurij Ružejnikov

Název práce:
Bayesovské odhadování pro adaptivní dynamické rozhodování

Autor: Jurij Ružejnikov

Studijní program: Aplikace přírodních věd
Obor: Matematické inženýrství
Druh práce: Bakalářská práce

Vedoucí práce: Dipl.-Eng. Tatiana V. Guy, PhD.
ÚTIA AVČR, v.v.i, Pod vodárenskou věží 4, 182 00 Praha 8
Konzultant: Ing. Marko Ruman
ÚTIA AVČR, v.v.i, Pod vodárenskou věží 4, 182 00 Praha

Abstrakt: V této práci vytvoříme algoritmus, který je schopen se chovat optimálně, vůči nějakému předem zadanému cíli. K tomu je třeba model systému, který je odhadnut pomocí bayesovkého odhadování, jehož hlavní výhodou je možnost aktualizace pravděpodobnostního modelu na základě nových dat. Pro řešení problému rozhodování jsme zvolili jednokrokovou iteraci hodnot stavů, která využívá odhadnutý model a poskytuje předpis pro volbu optimální akce. Navržené řešení bylo odhadnuto na simulovaných datech odrážejících problém online aukce v reálném čase.

Klíčová slova: aukce online reklam, chování v online burzách, zpětnovazební učení, bayesovké odhadování

Title:
Bayesian estimation for adaptive dynamic decision making

Author: Jurij Ružejnikov

Abstract: In this thesis, we create an algorithm that is able to behave optimally in respect to a predetermined goal. This requires a system model, which is learned via bayesian estimation, whose main advantage is the possibility of updating the probabilistic model based on new data. To solve the decision-making problem, we chose one-step value iteration, which uses the learned model to provide optimal action. The proposed approach was verified on simulated data describing real-time bidding problem.

Key words: real-time bidding, bidding behaviour, reinforcement learning, bayesian estimation

Obsah

Značení a termíny

$\mathbb{R}$ množina reálných čísel;

$\mathbb{N}$ množina přirozených čísel;

$\mathbb{N}_0$ množina přirozených čísel včetně nuly;

X diskrétní náhodná veličina;

$x \in \mathbb{R}$ realizace diskrétní náhodné veličiny X;

$p(Y)$ diskrétní pravděpodobnostní rozdělení náhodné veličiny Y;

$p(X, Y)$ sdružené pravděpodobnostní rozdělení náhodných veličin X a Y;

$p(X|Y)$ podmíněné pravděpodobnostní rozdělení X za podmínky Y;

$\theta \in \boldsymbol{\Theta}$ parametr modelu;

$t \in \boldsymbol{T}$ čas z diskrétní množiny $\boldsymbol{T}$, kde $\boldsymbol{T} \subset \mathbb{N}$;

$a_t \in \boldsymbol{A}$ akce v čase $t \in \boldsymbol{T}$ z diskrétní množiny akcí $\boldsymbol{A} \subset \mathbb{N}_0$;

$s_t \in \boldsymbol{S}$ stav v čase $t \in \boldsymbol{T}$ z diskrétní množiny stavů $\boldsymbol{S} \subset \mathbb{N}_0$;

$|\boldsymbol{A}| \in \mathbb{N}$ velikost diskrétní množiny akcí $\boldsymbol{A} \subset \mathbb{N}_0$;

$|\boldsymbol{S}| \in \mathbb{N}$ velikost diskrétní množiny stavů $\boldsymbol{S} \subset \mathbb{N}_0$;

$\pi(s_{t-1}) \in \boldsymbol{\Pi}$ strategie volby akce $a_t \in \boldsymbol{A}$ ve stavu $s_{t-1} \in \boldsymbol{S}$;

$\mathcal{B}$ borelovská σ-algebra v $\mathbb{R}$;

$r(s_t, a_t, s_{t-1})$ okamžitá odměna pro stav $s_t \in \boldsymbol{S}$, akci $a_t \in \boldsymbol{A}$ a stav $s_{t-1} \in \boldsymbol{S}$;

$\gamma \in \langle 0, 1 \rangle$ diskontní faktor, reprezentující preference agenta;

$v_\pi(s_t)$ hodnota stavu, při provedení strategie $\pi \in \Pi$

$b \in \mathbb{R}$ cena jednoho zobrazení reklamy;

Úvod

Co je RTB?

Reklamy jsou určeny k podpoření zájmu potenciálního zákazníka ke koupi jistého produktu nebo služby. Díky tomu, že v moderní době je internet nedílnou součástí každodenního života se přirozeně objevili i internetové reklamy. Velké společnosti, obvykle ty které vlastní webové prohlížeče, neustále zdokonalují sběr informací o uživatelích, aby jim mohli relevantně zobrazovat reklamu dle jejich zájmů. Tato interakce mezi uživatelem a inzerentem probíhá v prostředí online reklamní burzy [1]. Jedním z nejrychleji rostoucích přístupů k problému je nákup reklamy pomocí online aukcí v reálném čase (Real time bidding, dále jen RTB) [2]. Jedná se o aukci o jednotlivá místa na ploše prohlížeče uživatele, která proběhne v každém reklamním okně mezi jednotlivými inzerenty nabízejícími stejný produkt ještě před tím, než se uživateli stránka zobrazí. Výhodou tohoto přístupu pro inzerenty je možnost cílit reklamu na uživatele, kteří mají zájem o daný produktu či službu. Vydavatelé též optimalizují své zisky prodejem svých reklamních ploch více inzerentům současně.

Cílem této práce bude navrhnout a implementovat algoritmus, který je schopen přihazovat peněžní částku chytře, tedy maximalizovat počet uživatelů, kteří viděli nebo klikli na reklamu či koupili produkt/službu za co nejnižší cenu nakupených reklamních ploch.

Již existující řešení problematiky

Většina existujících řešení jsou založena na řešení problému optimálního příhozu jako statického optimalizačního problému (Stochastického gradientního posílení [3], Stromový log-normální model [3], lineární regrese [4]). Při obdržení nových dat se provádí celý odhad vždy znovu a zároveň se odhaduje pravděpodobnost výherní částky vložené do aukce. Tato řešení vedou k lepším ziskům než při volbě náhodného příhozu. Nicméně, RTB trh je vysoce dynamické prostředí. Rozdělení vytvořené na bázi statických dat se v takovém případě bude s nejvyšší pravděpodobností diametrálně lišit od skutečného rozdělení. Tedy v reálném čase je potřeba brát v potaz předešlou úspěšnost a neustále modifikovat odhad předpovědi výsledku aukce.

V první kapitole zavedeme základní pojmy a potřebnou teorii k řešení problému odhadování a rozhodování. V druhé kapitole definujeme problém online aukce reklam a zformulujeme řešení. V třetí kapitole popíšeme algoritmy potřebné pro odhadování a rozhodování, dále popíšeme způsob generování dat a simulaci pravděpodobnostního modelu. Ve čtvrté kapitole provedeme experimenty na generovaných datech a představíme výsledky navrženého řešení. V páté kapitole shrneme výsledky a prodiskutujeme budoucí vývoj.

Kapitola 1

Přehled použité teorie

1.1 Základní definice a vlastnosti pravděpodobnosti

V této kapitole zavedeme operace s pravděpodobnostními rozděleními, které jsou stěžejní v následujících kapitolách.

Definice 1.1 (Náhodná veličina)
Nechť: $(\mathbf{\Omega}, \boldsymbol{\mathcal{A}}, P)$ *je* **pravděpodobnostní prostor**, *kde*

- $\mathbf{\Omega}$ *je neprázdná diskrétní množina všech výsledků náhodného pokusu, neboli* **prostor elementárních jevů**. *Prvek* $\omega \in \mathbf{\Omega}$ *nazveme* **elementární jev**.
- $\boldsymbol{\mathcal{A}}$ *je množinový systém na* $\mathbf{\Omega}$, *který tvoří* σ*-algebru* [1]. *Prvek* $A \in \boldsymbol{\mathcal{A}}$ *nazveme* **náhodný jev**.
- $\mathbb{P}$ *je pravděpodobnostní míra na měřitelném prostoru* $(\mathbf{\Omega}, \boldsymbol{\mathcal{A}})$, $P : \boldsymbol{\mathcal{A}} \to \langle 0, 1 \rangle$. *Tuto míru budeme nadále označovat jako* **pravděpodobnost**.

Pak **náhodnou veličinou** *nazveme každé* $\mathcal{A}$*-měřitelné zobrazení* $X : (\mathbf{\Omega}, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$, *kde* $\mathcal{B}$, *je borelovská* $\sigma - algebra$, *pravě tehdy když* $(\forall B \in \mathcal{B})(X^{-1}(B) \in \mathcal{A})$.

Definice 1.2 (Realizace diskrétní náhodné veličiny)
Nechť $X : (\mathbf{\Omega}, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ *je diskrétní náhodná veličina (tj. když je obor hodnot nejvýše spočetná množina) definovaná na prostoru* $(\mathbf{\Omega}, \mathcal{A}, \mathbb{P})$, *pak* $X(\omega) = x$, *kde* x, *značí* **realizaci náhodné veličiny**.

Definice 1.3 (Diskrétní pravděpodobnostní rozdělení)
Nechť $X : (\mathbf{\Omega}, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ *je diskrétní náhodná veličina definovaná na* $(\mathbf{\Omega}, \mathcal{A}, \mathbb{P})$. *Nechť* x *je realizace náhodné veličiny* X. *Pak zobrazení* $p(X) : \mathcal{B} \to \langle 0, 1 \rangle$ *definované pro* X *jako*

$$p(X) = \mathbb{P}[\{\omega \in \mathbf{\Omega} : X(\omega) = x\}] \tag{1.1}$$

pro níž platí podmínka

$$\sum_{x} p(X = x) = 1 \tag{1.2}$$

[1] Pro podrobné matematické zavedení zavedení viz. [5]

nazýváme **diskrétní pravděpodobnostní rozdělení**.

V případě, že náhodná veličina je spojitá může být podobným způsobem zavedena hustota pravděpodobnosti [5].

Definice 1.4 (Sdružené pravděpodobnostní rozdělení)
Nechť $X, Y : (\mathbf{\Omega}, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ *jsou diskrétní náhodné veličiny definované na prostoru* $(\mathbf{\Omega}, \mathcal{A}, \mathbb{P})$. *Nechť* x *a* y *jsou realizace náhodných veličin* X *a* Y. *Pak* **sdruženým pravděpodobnostním rozdělením** *nazveme funkci*

$$p(X, Y) = \mathbb{P}[\{\omega \in \Omega : X(\omega) = x\} \cap \{\omega \in \Omega : Y(\omega) = y\}]. \tag{1.3}$$

Definice 1.5 (Marginální pravděpodobnostní rozdělení)
Nechť $X, Y : (\mathbf{\Omega}, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ *jsou diskrétní náhodné veličiny definované na prostoru* $(\mathbf{\Omega}, \mathcal{A}, \mathbb{P})$. *Nechť* $p(X, Y)$ *je sdružené pravděpodobnostní rozdělení* (1.3). *Nechť* x *a* y *jsou realizace náhodných veličin* X *a* Y. **Marginálním pravděpodobnostním rozdělením** *nazveme funkci*

$$p(X) = \sum_y p(X, Y = y). \tag{1.4}$$

Definice 1.6 (Podmíněné rozdělení pravděpodobnosti)
Nechť $X, Y : (\mathbf{\Omega}, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ *jsou náhodné veličiny definované na prostoru* $(\mathbf{\Omega}, \mathcal{A}, \mathbb{P})$ *a* x, y *jejich realizace. Mějme jisté, pevně dané,* y_0. *Nechť* $p(Y)$ *je diskrétní pravděpodobnostní rozdělení náhodné veličiny* Y (1.1). *Nechť* $p(X, Y)$ *je sdružené pravděpodobnostní rozdělení náhodných veličin* X *a* Y (1.3). *Nechť* $p(y_0) > 0$, *pak funkce*

$$\begin{aligned} p(X|Y = y_0) &= \mathbb{P}[\{\omega \in \Omega : X(\omega) = x\}|\{\omega \in \Omega : Y(\omega) = y_0\}] \\ &= \frac{\mathbb{P}[\{\omega \in \Omega : X(\omega) = x\}, \{\omega \in \Omega : Y(\omega) = y_0\}]}{p(\{\omega \in \Omega : Y(\omega) = y_0\})} \\ &= \frac{p(X, y_0)}{p(y_0)} \end{aligned} \tag{1.5}$$

je **podmíněné pravděpodobnostní rozdělení** *náhodné veličiny* X *za podmínky, že* $Y = y_0$, *pro které platí*

$$\sum_x p(X = x|Y = y_0) = 1, \forall y_0. \tag{1.6}$$

V praxi budeme používat dvě základní operace na pravděpodobnostních rozděleních:

- ***Řetězové pravidlo*** umožňuje konstruovat sdružené pravděpodobnostní rozdělení (1.3) pomocí podmíněného pravděpodobnostního rozdělení (1.5) následovně
$$p(X, Y) = p(X|Y)p(Y). \tag{1.7}$$
- ***Marginalizace*** umožňuje konstruovat podmíněné pravděpodobnosti s pomocí marginálního pravděpodobnostního rozdělení (1.4) následovně
$$p(X|Y) = \sum_z p(Z = z, X|Y). \tag{1.8}$$

Věta 1.7 (Bayesova věta)
Nechť $X, Y : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ jsou náhodné veličiny definované na prostoru $(\Omega, \mathcal{A}, \mathbb{P})$ a x, y jejich realizace. Nechť $p(X), p(Y) : \mathcal{B} \to \langle 0, 1\rangle$ jsou marginální rozdělení náhodných veličin X a Y (1.4). Nechť $p(X|Y)$ je podmíněné pravděpodobnostní rozdělení (1.5). Pak platí
$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)}, \tag{1.9}$$
kde $p(Y) > 0$.

Důkaz. Vyjádříme sdružené pravděpodobností rozdělení z řetězového pravidla (1.7) dvěma způsoby
$$p(X, Y) = p(X|Y)p(Y) = p(Y|X)p(X). \tag{1.10}$$
Z druhé rovnosti obdržíme vztah
$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)}. \tag{1.11}$$
□

Výše uvedené bylo odvozeno pro diskrétní případ. Vztahy (1.7) - (1.8) platí i pro spojitý případ (při záměně sumy za integrál v (1.8)). Pro rigorózní odvození viz. [5].

Pro zjednodušení zápisu nebudeme dále rozlišovat mezi náhodnou veličinou X a její realizací x. Z kontextu bude vždy jasné o kterou variantu se jedná.

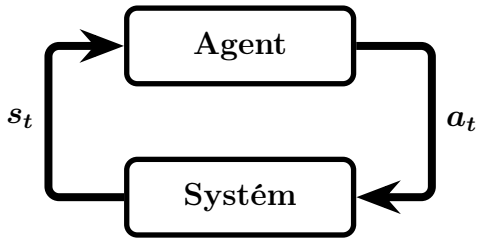
Pro jistá odvození bude také potřeba definovat Gamma funkci.

Definice 1.8 (Gamma funkce)
*Nechť $n \in \mathbb{N}$, pak **Gamma funkci** definujeme následujícím integrálem*
$$\Gamma(n) = \int_0^{+\infty} t^{n-1} e^t dt.$$

Dále platí
$$\forall n \in \mathbb{N} : \Gamma(n + 1) = n\Gamma(n). \tag{1.12}$$

1.2 Uzavřená smyčka agent-systém

Představme si příklad. Mějme autonomní vozidlo, které jede po vozovce. V každém časovém okamžiku vozidlo sleduje části vnějšího prostředí skrze své senzory. Snímá např.: přítomnost chodce, vzdálenost od okolních aut, typ a stav vozovky, dopravní značení atd. Poté na základě předem definovaných cílů, např.: dostat se bezpečně na jisté místo za nejkratší čas, vozidlo upravuje své chování, např.: změnu rychlosti či směru jízdy. V každém časovém okamžiku $t \in \boldsymbol{T}$ tedy vozidlo pozoruje část prostředí zvanou systém a na základě zpozorovaného stavu $s_{t-1} \in \boldsymbol{S}$ volí akci $a_t \in \boldsymbol{A}$ tak, aby dosáhlo svého (předem zadaného) cíle. Dále se systém dostává do stavu $s_t \in \boldsymbol{S}$ a tento proces se opakuje. Předchozí stav $s_{t-1} \in \boldsymbol{S}$ a akce $a_t \in \boldsymbol{A}$ mají tedy vliv na budoucí stav $s_t \in \boldsymbol{S}$. Viz. obrázek (1.1).
Naším cílem je popsat systém za účelem zlepšení předpovědi budoucího stavu $s_t \in \boldsymbol{S}$, aby agent na základě této předpovědi provedl optimální volbu akce $a_t \in \boldsymbol{A}$.
Klíčové k předpovědi budoucího stavu $s_t \in \boldsymbol{S}$ je znalost modelu systému, který lze získat na základě fyzikálních znalostí nebo z předem nasbíraných dat za předpokladu fixní struktury modelu. V další sekci se budeme zabývat odhadováním modelu systému na základě dostupných dat.



Obrázek 1.1: Rozhraní agent-systém

1.3 Bayesovské učení

Bayesovský přístup umožňuje aktualizovat své apriorní odhady skrze nová pozorování systému. Tedy vycházíme z apriorní informace a postupně ji doplňujeme o znalosti získávané z nově napozorovaných dat. Výsledek se používá pro další odhadování.

Uvažujme, že máme k dispozici data ve formě posloupností stavů a akcí do času $T \in \boldsymbol{T}$, tedy $(a_T, s_{T-1}, a_{T-1}, ..., s_1, a_1)$.

Definice 1.9
Náhodné veličiny uvažované v interakci agent-systém (viz. obrázek 1.1) lze popsat sdruženou pravděpodobností

$$p(s_T, a_T, s_{T-1}, a_{T-1}, ..., s_1, a_1). \tag{1.13}$$

Opakovanou aplikací řetězového pravidla (1.7) *dostaneme*

$$p(s_T, a_T, s_{T-1}, a_{T-1}, ..., s_1, a_1) = \prod_{t=2}^{T} p(s_t, a_t | s_{t-1}, a_{t-1}, ..., s_1, a_1) p(s_1, a_1). \tag{1.14}$$

A následnou aplikací řetězového pravidla (1.7) *na první člen* (1.14) *dostaneme*

$$\begin{aligned} &p(s_T, a_T, s_{T-1}, a_{T-1}, ..., s_1, a_1) \\ &\qquad = \prod_{t=2}^{T} p(s_t | a_t, s_{t-1}, a_{t-1}, ..., s_1, a_1) p(a_t | s_{t-1}, a_{t-1} ..., s_1, a_1) p(s_1, a_1), \end{aligned} \tag{1.15}$$

kde výraz

$$p(s_t | a_t, s_{t-1}, a_{t-1}, ..., s_1, a_1) \tag{1.16}$$

je ***model systému****, který vyjadřuje předpověď stavu $s_t \in \boldsymbol{S}$ v čase $t \in \boldsymbol{T}$ v závislosti na předešlých datech.*

Výraz

$$p(a_t | s_{t-1}, a_{t-1}, s_{t-1}, ..., s_1, a_1) \tag{1.17}$$

je ***rozhodovací pravidlo****, které vyjadřuje jak volit akce $a_t \in \boldsymbol{A}$ v čase $t \in \boldsymbol{T}$ na základě předešlých dat.*
A výraz

$$p(s_1, a_1) \tag{1.18}$$

vyjadřuje ***počáteční podmínky systému****, které nadále nebudeme explicitně vypisovat a zahrneme je do produktu následovně*

$$p(s_T, a_T, s_{T-1}, a_{T-1}, ..., s_1, a_1) = \prod_{t=1}^{T} p(s_t | a_t, s_{t-1}, a_{t-1}, ..., s_1, a_1) p(a_t | s_{t-1}, a_{t-1} ..., s_1, a_1). \tag{1.19}$$

Model systému (1.16) určuje závislost stavu systému na předchozích datech. Může se však stát, že nemáme k dispozici informace o všech veličinách ovlivňující tento stav.

Pro další odvození položíme následující předpoklad.

Předpoklad 1.10 (Markovský)
Současný stav $s_t \in \boldsymbol{S}$ závisí pouze na předchozím stavu $s_{t-1} \in \boldsymbol{S}$ a zvolené akci $a_t \in \boldsymbol{A}$.
Pro více informací viz. [6].

Po aplikaci Předpokladu 1.10 na (1.16) - (1.19) se vztahy zjednoduší na

$$p(s_T, a_T, s_{T-1}, a_{T-1}, ..., s_1, a_1) = \prod_{t=1}^{T} p(s_t|a_t, s_{t-1})p(a_t|s_{t-1}), \tag{1.20}$$

kde

$$p(s_t|a_t, s_{t-1}) \tag{1.21}$$

je model systému a

$$p(a_t|s_{t-1}) \tag{1.22}$$

je rozhodovací pravidlo.

Tvrzení 1.11
Předpokládejme, že známe model systému (1.21) *až na konečnou množinu parametrů $\theta \in \boldsymbol{\Theta}$, tj. tzv.* ***parametrický model systému***

$$p(s_t|a_t, s_{t-1}, \theta). \tag{1.23}$$

Předpověď stavu $s_t \in \boldsymbol{S}$ lze pak najít aplikací řetězového pravidla (1.7) a spojité verze marginalizačního pravidla (1.8) na (1.20), tedy

$$p(s_T, a_T, s_{T-1}, a_{T-1}, ..., s_1, a_1) = \prod_{t=1}^{T} \int_{\boldsymbol{\Theta}} p(s_t|a_t, s_{t-1}, \theta)p(a_t|s_{t-1}, \theta)p(\theta|s_{t-1}, a_{t-1}, ..., s_1, a_1)d\theta, \tag{1.24}$$

kde druhý člen není závislý na parametru $\theta \in \boldsymbol{\Theta}$ díky tzv. ***přirozeným podmínkám řízení*** [7], které odůvodněně předpokládají podmíněnou nezávislost akcí agenta na parametrech systému. Po aplikaci této podmínky na integrál (1.24) získává tvar

$$p(s_T, a_T, s_{T-1}, a_{T-1}, ..., s_1, a_1) = \prod_{t=1}^{T} \int_{\boldsymbol{\Theta}} p(s_t|a_t, s_{t-1}, \theta)p(a_t|s_{t-1})p(\theta|s_{t-1}, a_{t-1}, ..., s_1, a_1)d\theta, \tag{1.25}$$

kde $p(s_t|a_t, s_{t-1}, \theta)$ je parametrický model systému (1.23), $p(a_t|s_{t-1})$ je rozhodovací pravidlo (1.22) a

$$p(\theta|s_{t-1}, a_{t-1}, ..., s_1, a_1) \tag{1.26}$$

je model parametru $\theta \in \mathbf{\Theta}$.

Řešení úlohy předpovědi stavu vyžaduje následující kroky

1. Volba struktury parametrického modelu systému (1.23).
2. Odhad parametru modelu $\theta \in \mathbf{\Theta}$ (1.26).

V praxi mohou nastat dvě situace

1. Počet dostupných dat je fixovaný a lze jej zpracovat najednou, pak postačí odhadnou parametr $\theta \in \mathbf{\Theta}$ jen jednou.
2. Počet dat s časem narůstá a je třeba průběžně obnovovat znalost odhadu parametru $\theta \in \mathbf{\Theta}$ (1.26) na základě nových dat.

Věta 1.12 (Odhad parametru)
Pravděpodobnostní rozdělení popisující parametr systému $\theta \in \mathbf{\Theta}$ má tvar

$$p(\theta|s_{t-1}, a_{t-1}, ..., s_1, a_1) = \frac{\prod_{k=1}^{t-1} p(s_k|a_k, s_{k-1}, \theta)p(\theta)}{\int_{\mathbf{\Theta}} \prod_{k=1}^{t-1} p(s_k|a_k, s_{k-1}, \theta)p(\theta)d\theta}, \tag{1.27}$$

kde $p(\theta)$ reprezentuje apriorní rozdělení parametru $\theta \in \mathbf{\Theta}$.

Důkaz. Aplikujeme-li na aposteriorní distribuci parametru $\theta \in \mathbf{\Theta}$ (1.26) Bayesovu větu (1.9) dostaneme

$$p(\theta|s_{t-1}, a_{t-1}, ..., s_1, a_1) = \frac{p(s_{t-1}, a_{t-1}..., s_1, a_1|\theta)p(\theta)}{p(s_{t-1}, a_{t-1}..., s_1, a_1)}. \tag{1.28}$$

Dále aplikuji na jmenovatel (1.28) spojitou verzi marginálního pravidla (1.8) a řetězového pravidla (1.7)

$$p(\theta|s_{t-1}, a_{t-1}, ..., s_1, a_1) = \frac{p(s_{t-1}, a_{t-1}..., s_1, a_1|\theta)p(\theta)}{\int_{\mathbf{\Theta}} p(s_{t-1}, a_{t-1}..., s_1, a_1|\theta)p(\theta)d\theta}. \tag{1.29}$$

Aplikujeme opakovaně řetězové pravidlo (1.7)

$$p(\theta|s_{t-1}, a_{t-1}, ..., s_1, a_1) = \frac{\prod_{k=1}^{t-1} p(s_k, a_k|s_{k-1}, a_{k-1}..., s_1, a_1, \theta)p(\theta)}{\int_{\mathbf{\Theta}} \prod_{k=1}^{t-1} p(s_k, a_k|s_{k-1}, a_{k-1}..., s_1, a_1, \theta)p(\theta)d\theta}. \tag{1.30}$$

Využijeme ještě jednou řetězové pravidlo (1.7), přirozených podmínek řízení [7] a Předpoklad 1.10, potom

$$p(\theta|s_{t-1}, a_{t-1}, ..., s_1, a_1) = \frac{\prod_{k=1}^{t-1} p(s_k|a_k, s_{k-1}, \theta)p(a_k|s_{k-1})p(\theta)}{\int_{\mathbf{\Theta}} \prod_{k=1}^{t-1} p(s_k|a_k, s_{k-1}, \theta)p(a_k|s_{k-1})p(\theta)d\theta}. \tag{1.31}$$

A po vykrácení členu $p(a_k|s_{k-1})$ dostáváme kýžené tvrzení.

□

Značení 1.13
Pro zjednodušení zápisu zavedeme následující značení.
Trojice $(s, a, \overline{s})$ značí, že stav $\overline{s} \in \boldsymbol{S}$ a akce $a \in \boldsymbol{A}$ vede na stav $s \in \boldsymbol{S}$, tj. $(a, \overline{s}) \to s$. Pak pravděpodobnost, že se uskuteční přechod $(a, \overline{s}) \to s$ dle modelu $p(s|a, \overline{s}, \theta)$, který volíme jako neznámý parametr, je

$$\theta_{s|a,\overline{s}} = p(s|a, \overline{s}, \theta). \tag{1.32}$$

V diskrétním případě parametr $\theta \in \boldsymbol{\Theta}$ popisuje veškeré možné přechody $(a, \overline{s}) \to s$, $\forall s \in \boldsymbol{S}, a \in \boldsymbol{A}, \overline{s} \in \boldsymbol{S}$.

Pravděpodobnost (1.32) je právě veličina kterou budeme odhadovat.

Užitím značení (1.32) na model systému (1.23) pro časový přechod $(s_{k-1}, a_k) \to s_k$ dostáváme

$$p(s_k|a_k, s_{k-1}, \theta) = \theta_{s_k|a_k,s_{k-1}} = \prod_{s\in\boldsymbol{S},a\in\boldsymbol{A},\overline{s}\in\boldsymbol{S}} \theta_{s|a,\overline{s}}^{\delta(s,s_k)\delta(a,a_k)\delta(\overline{s},s_{k-1})}, \tag{1.33}$$

kde $\delta(i, j) = \begin{cases} 1, & \text{pokud } i = j, \\ 0, & \text{pokud } i \neq j. \end{cases}$ je Kroneckerovo delta funkce.

Pro libovolnou trojici $(s, a, \overline{s})$ můžeme zapsat statistiku popisující četnost jejich výskytu za uběhlých $t - 1 \in \boldsymbol{T}$ časových kroků jako

$$V_{t-1}(s, a, \overline{s}) = \sum_{k=1}^{t-1} \delta(s, s_k)\delta(a, a_k)\delta(\overline{s}, s_{k-1}). \tag{1.34}$$

Vztah (1.34) se dá zapsat rekurzivně následujícím způsobem

$$V_t(s, a, \overline{s}) = \delta(s, s_t)\delta(a, a_t)\delta(\overline{s}, s_{t-1}) + V_{t-1}(s, a, \overline{s}), \tag{1.35}$$

jímž po každém novém pozorování (s_t, a_t, s_{t-1}) aktualizujeme statistiku V_{t-1} na V_t.

Člen $p(\theta)$ ve vztahu pro odhad parametru (1.27) popisuje apriorní představu o možných přechodech mezi jednotlivými stavy $s \in \boldsymbol{S}$ za podmínky konkrétní akce $a \in \boldsymbol{A}$ a předchozího stavu $\overline{s} \in \boldsymbol{S}$. Pro naší úlohu ho zvolíme v následujícím tvaru

$$p(\theta) \propto \prod_{s\in\boldsymbol{S},a\in\boldsymbol{A},\overline{s}\in\boldsymbol{S}} \theta_{s|a,\overline{s}}^{V_0(s,a,\overline{s})-1}, \tag{1.36}$$

kde $V_0(s, a, \overline{s}) > 0$ je apriorní předpoklad počtu výskytu trojic $(s, a, \overline{s})$.

Nyní dosadíme vztah pro relativní četnosti výskytu parametru $\theta_{s|a,\overline{s}} \in \boldsymbol{\Theta}$ (1.34), volbu parametrického rozdělení (1.33) a volbu apriorního rozdělení parametru θ (1.36) do pravděpodobnostního rozdělení parametru θ (1.27) a dostaneme

$$p(\theta|a_{t-1}, s_{t-1}, ..., a_1, s_1) = \frac{\prod_{s\in\boldsymbol{S},a\in\boldsymbol{A},\overline{s}\in\boldsymbol{S}} \theta_{s|a,\overline{s}}^{V_{t-1}(s,a,\overline{s})}\theta_{s|a,\overline{s}}^{V_0(s,a,\overline{s})-1}}{\int_{\boldsymbol{\Theta}} \prod_{s\in\boldsymbol{S},a\in\boldsymbol{A},\overline{s}\in\boldsymbol{S}} \theta_{s|a,\overline{s}}^{V_{t-1}(s,a,\overline{s})}\theta_{s|a,\overline{s}}^{V_0(s,a,\overline{s})-1} d\theta}. \tag{1.37}$$

Vztah (1.37) upravíme na

$$p(\theta|a_{t-1}, s_{t-1}, ..., a_1, s_1) = \frac{\prod_{s\in\boldsymbol{S}, a\in\boldsymbol{A}, \overline{s}\in\boldsymbol{S}} \theta_{s|a,\overline{s}}^{V_{t-1}(s,a,\overline{s})+V_0(s,a,\overline{s})-1}}{\int_{\boldsymbol{\Theta}} \prod_{s\in\boldsymbol{S}, a\in\boldsymbol{A}, \overline{s}\in\boldsymbol{S}} \theta_{s|a,\overline{s}}^{V_{t-1}(s,a,\overline{s})+V_0(s,a,\overline{s})-1} d\theta}. \tag{1.38}$$

Pro zkrácení zápisu a přehlednost označíme

$$R_{t-1}(s, a, \overline{s}) = V_{t-1}(s, a, \overline{s}) + V_0(s, a, \overline{s}), \tag{1.39}$$

což vyjadřuje četnost výskytu trojic $(s, a, \overline{s})$ do času $t-1 \in \boldsymbol{T}$ včetně apriorního předpokladu jejich výskytu.

Užitím značení (1.39) na pravděpodobnostní rozdělení parametru $\theta \in \boldsymbol{\Theta}$ (1.38) dostaneme

$$p(\theta|a_{t-1}, s_{t-1}, ..., a_1, s_1) = \frac{\prod_{s\in\boldsymbol{S}, a\in\boldsymbol{A}, \overline{s}\in\boldsymbol{S}} \theta_{s|a,\overline{s}}^{R_{t-1}(s,a,\overline{s})-1}}{\int_{\boldsymbol{\Theta}} \prod_{s\in\boldsymbol{S}, a\in\boldsymbol{A}, \overline{s}\in\boldsymbol{S}} \theta_{s|a,\overline{s}}^{R_{t-1}(s,a,\overline{s})-1} d\theta}. \tag{1.40}$$

Nalezli jsme tedy vztah pro odhad parametru $\theta \in \boldsymbol{\Theta}$ a jedná se o Dirichletovu distribuci [8].

Před dalšími úpravami si připomeneme vztah pro interakci agent-systém (1.13) upraveného do tvaru s parametrem θ (1.25) a tedy

$$\begin{aligned} &p(s_T, a_T, s_{T-1}, a_{T-1}, ..., s_1, a_1) \\ &\qquad = \prod_{t=1}^{T} \int_{\boldsymbol{\Theta}} p(s_t|a_t, s_{t-1}, \theta) p(a_t|s_{t-1}, \theta) p(\theta|s_{t-1}, a_{t-1}, ..., s_1, a_1) d\theta, \end{aligned} \tag{1.41}$$

který upravíme do tvaru vyjadřující přirozené podmínky řízení [7]

$$\begin{aligned} &p(s_T, a_T, s_{T-1}, a_{T-1}, ..., s_1, a_1) \\ &\qquad = \prod_{t=1}^{T} p(a_t|s_{t-1}) \int_{\boldsymbol{\Theta}} p(s_t|a_t, s_{t-1}, \theta) p(\theta|s_{t-1}, a_{t-1}, ..., s_1, a_1) d\theta. \end{aligned} \tag{1.42}$$

A dále se soustředíme pouze na integrál v (1.42), který reprezentuje model systému (1.21), tedy

$$p(s_t|a_t, s_{t-1}) = \int_{\boldsymbol{\Theta}} p(s_t|a_t, s_{t-1}, \theta) p(\theta|s_{t-1}, a_{t-1}, ..., s_1, a_1) d\theta. \tag{1.43}$$

Věta 1.14 (Model systému)
Vztah (1.43) *může být vyčíslen v následujícím tvaru*

$$p(s_t|a_t, s_{t-1}, R_{t-1}) = \frac{R_{t-1}(s_t, a_t, s_{t-1})}{\sum_{s\in\boldsymbol{S}} R_{t-1}(s, a_t, s_{t-1})}, \tag{1.44}$$

kde $R_{t-1}(s_t, a_t, s_{t-1})$ *je značení* (1.39).

Důkaz. Dosadíme konkrétní volbu pravděpodobnostního rozdělení parametru $\theta \in \boldsymbol{\Theta}$ (1.40) do (1.43)

$$p(s_t|a_t, s_{t-1}, R_{t-1}) = \int_{\boldsymbol{\Theta}} p(s_t|a_t, s_{t-1}, \theta) \frac{\prod_{s\in\boldsymbol{S},a\in\boldsymbol{A},\overline{s}\in\boldsymbol{S}} \theta_{s|a,\overline{s}}^{R_{t-1}(s,a,\overline{s})-1}}{\int_{\boldsymbol{\Theta}} \prod_{s\in\boldsymbol{S},a\in\boldsymbol{A},\overline{s}\in\boldsymbol{S}} \theta_{s|a,\overline{s}}^{R_{t-1}(s,a,\overline{s})-1} d\theta} d\theta. \quad (1.45)$$

Dále dosadíme zvolený parametrický model systému (1.33) do (1.45), dostáváme

$$p(s_t|a_t, s_{t-1}, R_{t-1}) = \int_{\boldsymbol{\Theta}} \prod_{s\in\boldsymbol{S},a\in\boldsymbol{A},\overline{s}\in\boldsymbol{S}} \theta_{s|a,\overline{s}}^{\delta(s,s_t)\delta(a,a_t)\delta(\overline{s},s_{t-1})} \times \frac{\prod_{s\in\boldsymbol{S},a\in\boldsymbol{A},\overline{s}\in\boldsymbol{S}} \theta_{s|a,\overline{s}}^{R_{t-1}(s,a,\overline{s})-1}}{\int_{\boldsymbol{\Theta}} \prod_{s\in\boldsymbol{S},a\in\boldsymbol{A},\overline{s}\in\boldsymbol{S}} \theta_{s|a,\overline{s}}^{R_{t-1}(s,a,\overline{s})-1} d\theta} d\theta. \quad (1.46)$$

Nyní je třeba si všimnout, že se jedná o střední hodnotu Dirichletovi distribuce, která má následující tvar [8]

$$\mathbb{E}[\theta_{s_t|a_t,s_{t-1}}|s_t, a_t, s_{t-1}, R_{t-1}] = p(s_t|a_t, s_{t-1}, R_{t-1}) = \frac{R_{t-1}(s_t, a_t, s_{t-1})}{\sum_{s\in\boldsymbol{S}} R_{t-1}(s, a_t, s_{t-1})}. \quad (1.47)$$

□

1.4 Markovské rozhodovací procesy

Při interakci agenta se systémem má agent určitý cíl vůči systému. Nechť agent pozoruje stav $s_{t-1} \in \boldsymbol{S}$ v jistém čase $t-1 \in \boldsymbol{T}$. Jeho rozhodnutí jsou omezená na výběr akcí $a_t \in \boldsymbol{A}$ v čase $t \in \boldsymbol{T}$, kterými je schopen ovlivňovat systém. Zároveň předpokládejme, že množina všech časových okamžiků $\boldsymbol{T}$ je spočetná.
Kdyby měl agent k dispozici informaci o stavu $s_t \in \boldsymbol{S}$, tj. neměl by potřebu provádět odhad, pak by volba akce $a_t \in \boldsymbol{A}$ nebyla problém a agent by vždy volil takovou, která mu přinese největší okamžitou odměnu.

Definice 1.15 (Okamžitá odměna)
Zobrazení $r : \boldsymbol{S} \times \boldsymbol{A} \times \boldsymbol{S} \to \mathbb{R}$, které je shora omezené a reprezentuje preference agenta, označíme jako

$$r(s_t, a_t, s_{t-1}), \quad (1.48)$$

$\forall s_t \in \boldsymbol{S}, a_t \in \boldsymbol{A}, s_{t-1} \in \boldsymbol{S}$ *a nazveme jej* ***okamžitou odměnou****.*

Definice 1.16 (Strategie)
Zobrazení $\pi : \boldsymbol{S} \to \boldsymbol{A}$, reprezentující předpis pro volbu akce $a \in \boldsymbol{A}$ pro každý stav $s \in \boldsymbol{S}$ rozumíme jako ***strategie*** *a označíme*

$$\pi(s). \quad (1.49)$$

Množinu všech strategií π označíme $\boldsymbol{\Pi}$.

Obecně je strategie tvořena časovou posloupností (1.49). Jelikož se v této práci zabýváme stacionárním případem, můžeme využít tuto zjednodušující definici.

Racionální agent by měl vždy volit akci $a_t \in \boldsymbol{A}$, která mu přinese maximální odměnu (tzv. princip maximalizace užitku [9]). Ovšem díky neurčitosti se dozvíme hodnotu odměny až po provedení akce $a_t \in \boldsymbol{A}$. Proto využijeme očekávanou odměnu.

Definice 1.17 (Očekávaná odměna)
*Zobrazení $r : \boldsymbol{A} \times \boldsymbol{S} \to \mathbb{R}$, které reprezentuje **očekávanou odměnu** ve stavu $s_t \in \boldsymbol{S}$ definujeme jako*

$$r(a_t, s_{t-1}) = \sum_{s \in \boldsymbol{S}} r(s, a_t, s_{t-1}) p(s|a_t, s_{t-1}), \tag{1.50}$$

kde $p(s|a_t, s_{t-1})$ je model systému (1.21).

Cílem agenta je nyní najít strategii maximalizující celkovou očekávanou odměnu.

Definice 1.18 (Hodnota stavu)
*Funkce $v_\pi : \boldsymbol{S} \to \mathbb{R}$, která reprezentuje **hodnotu stavu** $s \in \boldsymbol{S}$ při provedení strategie $\pi \in \boldsymbol{\Pi}$ a rovná se součtu očekávané okamžité odměny a budoucí diskontní hodnotu stavu*

$$v_\pi(s) = \sum_{s \in \boldsymbol{S}} p(s|a_t, s_{t-1})(r(s, a_t, s_{t-1}) + \gamma v_\pi(s)), \tag{1.51}$$

*kde $p(s|a_t, s_{t-1})$ je pravděpodobnostní model (1.21), $r(s, a_t, s_{t-1})$ je okamžitá odměna (1.48) a $\gamma \in (0, 1)$ je **diskontní faktor**, který reprezentuje preferenci agenta a kde hodnoty blíže k 0 reprezentují preferenci krátkodobé odměny a hodnoty blíže k 1 reprezentují preferenci dlouhodobé odměny.*

Rovnici (1.51) nazveme **Bellmanova rovnice**.

Výsledkem volby akce $a_t \in \boldsymbol{A}$ v čase $t \in \boldsymbol{T}$ je tedy:

1. Přechod do stavu $s_t \in \boldsymbol{S}$, který je určen modelem (1.21).
2. Agent obdrží okamžitou odměnu (1.48) tj. $r(s_t, a_t, s_{t-1})$.

Množinu objektů

$$\{\boldsymbol{T}, \boldsymbol{S}, \boldsymbol{A}, p(s_t|a_t, s_{t-1}), r, \gamma\} \tag{1.52}$$

budeme nazývat ***Markovský rozhodovací proces***, kde $p(s_t|a_t, s_{t-1})$ je model (1.21).

Naším cílem je najít optimální akci $a_t \in \boldsymbol{A}$, která přinese maximální odměnu.

Definice 1.19 (Optimální akce)
Takovou akci $a_t^* \in \boldsymbol{A}$ pro kterou platí

$$a_t^* = \underset{a_t \in \boldsymbol{A}}{argmax} \sum_{s \in \boldsymbol{S}} p_t(s|a_t, s_{t-1})(r(s, a_t, s_{t-1}) + \gamma v_\pi(s)), \tag{1.53}$$

*nazveme **optimální akce**, kde $p(s|a_t, s_{t-1})$ je model (1.21), $r(s, a_t, s_{t-1})$ je okamžitá odměna (1.48) a $v_\pi(s)$ je hodnota stavu (1.51).*

Hodnotu stavu při optimální akci $a_t^* \in \boldsymbol{A}$ ve stavu $s_{t-1} \in \boldsymbol{S}$ budeme nazývat ***optimální hodnota stavu*** a značit

$$v_*(s_{t-1}) = \max_{a_t \in \boldsymbol{A}} v_\pi(s_{t-1}), \forall s_{t-1} \in \boldsymbol{S}. \tag{1.54}$$

1.5 Dynamické programování

Dynamické programování je algoritmus, který hledá optimální strategii $\pi \in \boldsymbol{\Pi}$. Stěžejním prvkem algoritmu bude Bellmanova rovnice optimality (1.51), přesněji řečeno její iterační varianta. Cílem této sekce bude převést tyto rovnice do iterativních pravidel, které konvergují k jejich optimální hodnotě.

Nejprve s užitím vztahu (1.51) spočteme $v_\pi(s_{t-1})$ pro všechna $s_{t-1} \in \boldsymbol{S}$ a libovolnou strategii $\pi \in \boldsymbol{\Pi}$. Označme $|\boldsymbol{S}| \in \mathbb{N}_0$ počet prvků množiny $\boldsymbol{S}$. Jedná se tedy o $|\boldsymbol{S}|$ rovnic o $|\boldsymbol{S}|$ neznámých. Řešení takové soustavy je zaručeno a je jedinečné [10]. Využijeme iteračních metod k nalezení řešení. Předpokládejme tedy počáteční hodnoty stavů pro všechny stavy $s_{t-1} \in \boldsymbol{S}$, kde $v_0(s_{t-1}) \in \mathbb{R}$. Postupnou iterací pro každý stav $s_{t-1} \in \boldsymbol{S}$ budeme konvergovat ke skutečné hodnotě $v_\pi(s_{t-1})$ s užitím (1.51) jako iterativního obnovovacího pravidla

$$v_{k+1}(s_{t-1}) = \sum_{s \in \boldsymbol{S}} p(s|a_t, s_{t-1})(r(s, a_t, s_{t-1}) + \gamma v_k(s)). \tag{1.55}$$

Lze prokázat, že v_k konverguje k v^* [10].
Kritérium konvergence je pak

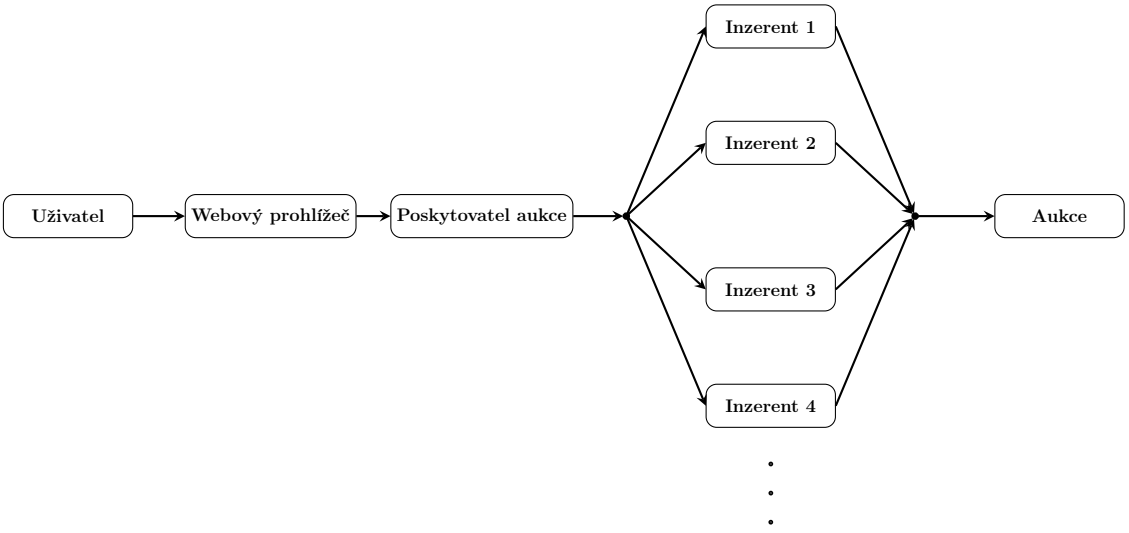
$$|v_{k+1}(s_{t-1}) - v_k(s_{t-1})| < \frac{\epsilon(1-\gamma)}{\gamma}, \tag{1.56}$$

kde $\epsilon \in \mathbb{R}$ je maximální dovolená chyba odhadu hodnoty stavu.
Tento algoritmus se nazývá ***iterace hodnot***. Potom, co najdeme optimální hodnotu stavů (1.51) pro všechny stavy $s_{t-1} \in \boldsymbol{S}$, najdeme pomocí (1.53) optimální akci $a_t \in \boldsymbol{A}$.

Kapitola 2

Formulace RTB jako rozhodovací úlohy

Představme si, že si prohlížíme stránky na internetu. Je dobrá šance, že narazíme na stránku, na které jsou vyhrazena místa pro reklamu. Pokaždé, když probíhá načítání takové stránky v našem prohlížeči, probíhá na každém z těchto reklamních míst aukce následujícím způsobem.
Takzvaný provozovatel aukce vybere účastníky aukce (dále jen ***inzerenty***) dle našich preferencí, které má k dispozici z naší historie surfování na webu, tedy např.: určí, že máme zájem o úvěr od a vybere tedy do aukce různé banky. Každý inzerent následně přihazuje do aukce a soutěží o reklamní pozici. Vítěz obdrží právo zobrazit svojí reklamu a ostatní inzerenti o vložené peníze přicházejí. Aukce se přesouvá na další reklamní pozici na stránce a proces začíná znovu. Po načtení stránky se vám zobrazí reklamy jednotlivých vítězů. Tento proces je vyobrazen na Obrázku 2.1. Poskytovatel aukce vždy vybírá inzerenty stejného zaměření. Například určí, na základě našich preferencí, že máme zájem o hypotéku, přidělí do aukce o reklamní pole na námi zobrazované stránce pouze banky. Tento fakt dovoluje předpokládat, že úloha je dynamická a aukce probíhají opakovaně s inzerenty, kteří se snaží prodat stejný produkt a u stejného poskytovatele a lze tedy tvrdit, že z opakované účasti na aukci lze načerpat zkušenosti.

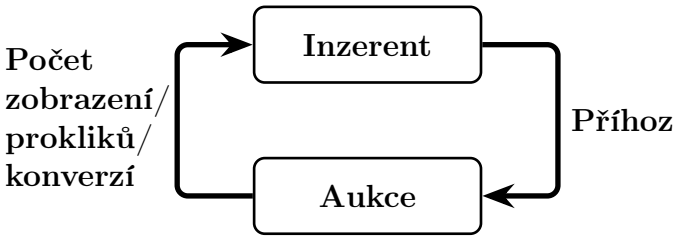


Obrázek 2.1: Online aukce

V této práci se stavíme do pozice inzerenta a naším cílem je navrhnout chytrého agenta, který bude optimalizovat přihazování do aukce.

2.1 Aukce jako rozhodování v uzavřené smyčce

Postavme se nyní do pozice jednoho z inzerentů na Obrázku 2.1. Situaci budeme modelovat jako uzavřenou smyčku agent-systém ze Sekce 1.2.
Obecně je cílem inzerenta získat co nejvíce zobrazení reklamy, nejvíce prokliků reklamy a následného zakoupení služby uživatelem (dále jen konverze) v každém čase $t \in \boldsymbol{T}$ [11]. Neboli, dále v řeči Sekce 1.2, dosáhnout co největší hodnoty stavu $s_t \in \boldsymbol{S}$. Inzerent, dále agent, volí v čase $t \in \boldsymbol{T}$ akci $a_t \in \boldsymbol{A}$ ve formě přihození do aukce, která následně (typicky se zpožděním minut, hodin nebo dní) vrací informaci o počtu zobrazení/prokliků/konverzí. Proces je vyobrazen na Obrázku 2.2.



Obrázek 2.2: Aukce jako uzavřená smyčka agent-systém

Systém viz. obrázek 1.1 je tvořen aukcí která se skládá z poskytovatele a ostatních inzerentů. Značení ze Sekce 1.2 tedy nabývá následujícího tvaru

- Poskytovatel aukce a ostatní inzerenti - ***systém***

- Náš inzerent - ***agent***
- Časový okamžik $t \in \boldsymbol{T}$, množina časových okamžiků $\mathbf{T}$
- Příhoz do aukce v čase $t \in \boldsymbol{T}$ - akce $a_t \in \boldsymbol{A}$
- Počet zobrazení v čase $t \in \boldsymbol{T}$ - stav $s_t \in \boldsymbol{S}$

Úloha optimalizace příhozu má jisté problémy

- Data jsou těžko dostupná, jelikož podléhají firemnímu tajemství a proto nejsou k dispozici z veřejných zdrojů.
- Dat je velké množství a jsou k dispozici se zpožděním. Data počtů zobrazení/prokliků/konverzí jsou často dostupná až na konci dne.
- Data jsou diskrétní v hodnotách stavů $s_t \in \boldsymbol{S}$, spojitá v hodnotách akcí $a_t \in \boldsymbol{A}$ a času $t \in \boldsymbol{T}$.
- Data jsou taktéž závislá na externích parametrech např.: Lepší počasí má za následek menší návštěvnost webových stránek. Roční období má vliv na čas uživatele strávený na webu. Nečekané události mohou ovlivnit poptávku po inzerovaném produktu atd.

2.2 Zjednodušující předpoklady

Vzhledem k problémům popsaným v Sekci 2.1 nyní provedeme jistá zjednodušení pro naši implementaci.

Předpoklad 2.1

- *Cílem inzerenta je maximalizovat pouze počet zobrazení reklamy jakožto jednorozměrný stav $s_t \in \boldsymbol{S}$.*
- *Informace o počtu zobrazení je dostupná s libovolným zpožděním.*
- *Data jsou diskrétní v hodnotách stavů $s_t \in \boldsymbol{S}$, v hodnotách akcí $a_t \in \boldsymbol{A}$ i v hodnotách času $t \in \boldsymbol{T}$.*

Poznámka 2.2
Zahrnutí počtu prokliků nebo konverzí služby do stavu $s_t \in \boldsymbol{S}$ není těžké. Stav může být zadefinován jako vektor ve formě (počet zobrazení, počet prokliků, počet konverzí), kde jednotlivé položky jsou pozorovatelné. Informace o prokliku na reklamu je lehce pozorovatelná, ale údaj o případné konverzi není přímo pozorovatelný, jelikož je pro každého uživatele pozorován jako kliknutí na jistou část reklamní stránky, např.: tlačítko "Koupit", ale uživatel si to na další stránce může před placením rozmyslet.

2.3 Agent pro optimalizaci příhozu

Navržený algoritmus se skládá z několika částí.

1. Simulace chování aukce
   Zde byl vytvořen model chování aukce (systém), který generuje počet zobrazení (nový stav) na základě navrženého příhozu (akce) a minulého stavu. Tento model využíváme pro imitaci skutečné aukce.

2. Odhadování modelu aukce, viz. Sekce 1.3
   Zde byl navržen algoritmus realizující bayesovské odhadování. Generovaná data na základě simulovaného modelu budou použita pro odhadování aproximovaného modelu systému (1.44), který následně využijeme v metodě iterace hodnot.

3. Návrh optimální akce
   Zde byl vytvořen algoritmus realizující metodu iterace hodnot, viz. Sekce 1.5. Spočtená optimální akce je aplikována v simulaci pro vygenerování nového stavu.

Detaily implementace jsou popsané v Kapitole 3.
K nalezení optimální strategie, je třeba pro rovnici (1.53) definovat tvar okamžité odměny (1.48), která vyjadřuje preference agenta.
Vzhledem k cíli agenta (viz. Předpoklad 2.1) je okamžitá odměna navržena ve tvaru

$$r(s_t, a_t, s_{t-1}) = s_t \cdot b - a_t, \tag{2.1}$$

kde $b \in \mathbb{R}^+$ je cena za jedno zobrazení v CZK, $s_t \in \boldsymbol{S}$ je počet zobrazení a $a_t \in \boldsymbol{A}$ je příhoz do aukce.
Obor hodnot této funkce reprezentuje čistý zisk v CZK.

Kapitola 3

Implementace algoritmů

Pro realizaci řešení bylo zvoleno prostředí Matlab [12].
Simulovaná data jsou stavy $s_t \in \boldsymbol{S}$ a akce $a_t \in \boldsymbol{A}$, v jistých časových okamžicích $t \in \boldsymbol{T}$, které ukládáme do příslušných 1D polí.

3.1 Simulace

Model systému pro simulaci je vytvořen následujícím způsobem.
Vygenerujeme 3D pole, rozměru $\boldsymbol{S} \times \boldsymbol{A} \times \boldsymbol{S}$, pravděpodobností přechodu $p(s_t|a_t, s_{t-1})$ pro všechna $s_t \in \boldsymbol{S}$, $a_t \in \boldsymbol{A}$ a $s_{t-1} \in \boldsymbol{S}$, tak aby největší pravděpodobnost ležela na diagonále jednotlivých matic příslušných ke všem $s_{t-1} \in \boldsymbol{S}$ a ostatní hodnoty ve sloupci klesaly dle normálního rozdělení. Přičemž pro každou matici příslušnou k $s_{t-1} \in \boldsymbol{S}$ platí, že stavy $s_t \in \boldsymbol{S}$ a akce $a_t \in \boldsymbol{A}$ jsou uspořádány od nejmenší po největší. Tato volba vyplývá z předpokladu, že při volbě akce $a_t \in \boldsymbol{A}$ příslušné k diagonálnímu členu a příslušného stavu $s_t \in \boldsymbol{S}$, jehož pozice vzhledem k jeho velikosti vůči ostatním stavům $s_t \in \boldsymbol{S}$, je stejná jako pozice akce $a_t \in \boldsymbol{A}$ vzhledem k velikosti ostatních akcí. Má v tomto místě vyšší pravděpodobnost přechodu do stavu $s_t \in \boldsymbol{S}$ než u akce $a_t \in \boldsymbol{A}$, která je velikostně menší či větší.

Příklad 3.1
Mějme stavový prostor S = [300,200,100] a prostor akcí A = [30, 20, 10]. Pro volbu akce $a_t = 20$ *dává smysl předpokládat, že se uskuteční přechod do stavu* $s_t = 20$ *s větší pravděpodobností než do stavu* $s_t = 30$ *či* $s_t = 10$. *Stejně tak pro volbu akce* $a_t = 30$ *má smysl předpokládat, že přechod do stavu* $s_t = 30$ *má větší pravděpodobnost než přechod do stavu* $s_t = 20$ *a do stavu* $s_t = 10$ *ještě nižší atd.*

Následně aplikujeme na každou hodnotu pravděpodobnosti přechodu $p(s_t|a_t, s_{t-1})$ náhodný šum, tím se maximální hodnota pravděpodobnosti na některých místech pohne z diagonály, což se dá interpretovat jako vliv externích nepozorovaných proměnných na model. Pro možnost opětovného vygenerování stejných výsledků je v náhodném generování modelu, při aplikování šumu, využíván seed. Toto řešení je neimplementováno pro podobně velké velikosti prostoru stavů a akcí tj. $|\boldsymbol{S}| \approx |\boldsymbol{A}|$,

kde $\boldsymbol{S}, \boldsymbol{A} \in \mathbb{N}_0$ a je třeba je takto volit, jinak není zaručeno ani přibližné dodržení vzoru popsaného výše. Pro ilustrační účely úlohy je však postačující.

3.2 Odhadování

Alokujeme dvě 3D pole rozměru $\boldsymbol{S} \times \boldsymbol{A} \times \boldsymbol{S}$. Jedno pro sledování počtu výskytů trojic (1.39) $r(s_t, a_t, s_{t-1})$, které naplníme jedničkami a druhé pro sledování počtu výskytů dvojic $\sum_{s \in \boldsymbol{S}} r(s, a_t, s_{t-1})$, které naplníme hodnotou $|\boldsymbol{S}| \in \mathbb{N}$ reprezentující velikost stavového prostoru $\boldsymbol{S} \subset \mathbb{N}_0$. Následně vygenerujeme 3D pole stejných rozměrů, reprezentující pravděpodobnosti přechodu $p(s_t|a_t, s_{t-1})$ a naplníme jej hodnotami dle (1.44).
Po vygenerování nového stavu $s_t \in \boldsymbol{S}$ v čase $t \in \boldsymbol{T}$, zvýšíme počet výskytu trojic do času $t \in \boldsymbol{T}$ (1.39) $r(s_t, a_t, s_{t-1})$ o 1, taktéž zvýšíme výskyt dvojic $\sum_{s \in \boldsymbol{S}} r(s, a_t, s_{t-1})$ o 1 a s pomocí (1.44) přepočteme model $p(s_t|a_t, s_{t-1})$. Celý proces je popsán v Algoritmu 1.

Algorithm 1 Odhadování modelu

1: **Inicializace**:
2: *tabulka_trojic* = 3D pole jedniček, *tabulka_dvojic* = 3D pole hodnot $|S| \in \mathbb{N}_0$, *predikovaný_model* = 3D pole hodnot podílu. $\frac{1}{|S|}$, každé pole je rozměru $\boldsymbol{S} \times \boldsymbol{A} \times \boldsymbol{S}$
3: **for each** $t \in \boldsymbol{T}$ **do**
4: $\quad$ *tabulka_trojic*$(s_t, a_t, s_{t-1}) \leftarrow$ *tabulka_trojic*$(s_t, a_t, s_{t-1}) + 1$
5: $\quad$ *tabulka_dvojic*$(s_t, a_t, s_{t-1}) \leftarrow$ *tabulka_dvojic*$(s_t, a_t, s_{t-1}) + 1$
6: $\quad$ $p(s_t|a_t, s_{t-1}) = \frac{\textit{tabulka\_trojic}(s_t, a_t, s_{t-1})}{\textit{tabulka\_dvojic}(s_t, a_t, s_{t-1})}$
7: **end for**
8: Výstup: Odhadnutý model systému.

3.3 Rozhodování

Návrh optimální akce $a_t \in \boldsymbol{A}$ v čase $t \in \boldsymbol{T}$ je prováděn s využitím iterativního obnovovacího pravidla (1.55), s parametrem $\gamma = 0$. Což dle definice (1.18), reprezentuje preferenci krátkodobé odměny. Jedná se o jednokrokovou optimalizaci.
Nejprve inicializujeme hodnoty všech stavů (1.51) v každém čase $t \in \boldsymbol{T}$ pro všechna $s_t \in \boldsymbol{S}$ na $v(s_t) = 0$. Okamžitou odměnu (1.48) $r(s, a_r, s_{t-1})$ pro všechna $s \in \boldsymbol{S}$, $a_t \in \boldsymbol{A}$, $s_{t-1} \in \boldsymbol{S}$ budeme počítat dle předpisu v Sekci 2.3. Jelikož neznáme model systému (1.21), budeme v (1.55) využívat odhadnutý model (1.44), který aktualizujeme v každém časovém kroku na základě nových dat. Výsledky pro každý model porovnáme v Kapitole 4.
V každém časovém kroku $t \in \boldsymbol{T}$ spočteme maximální hodnotu stavu (1.55) a odpovídající optimální akci $a_t \in \boldsymbol{A}$. Celý proces je popsán v algoritmu 2.

Algorithm 2 Návrh optimální akce $a_t \in \boldsymbol{A}$

1: **Inicializace**:
2: $v(s_t) = 0$ pro každé $s_t \in \boldsymbol{S}$, $t \in \boldsymbol{T}$, okamžitá odměna $r(s, a_t, s_{t-1})$ pro všechna $s \in \boldsymbol{S}$, $a_t \in \boldsymbol{A}$, $s_{t-1} \in \boldsymbol{S}$ a $\gamma = 0$.
3: **for each** $t \in \boldsymbol{T}$ **do**
4: $\quad v(s_t) \leftarrow \max_{a_t \in \boldsymbol{A}} \sum_{s \in \boldsymbol{S}} p(s|a_t, s_{t-1}) r(s, a_t, s_{t-1})$
5: $\quad a_t^* \leftarrow \underset{a_t \in \boldsymbol{A}}{argmax} \sum_{s \in \boldsymbol{S}} p(s|a_t, s_{t-1}) r(s, a_t, s_{t-1})$
6: **end for**
7: Výstup: optimální akce $a_t \in \boldsymbol{A}$ v čase $t \in \boldsymbol{T}$.

3.4 Generování dat

Data jsou generována dle modelu vytvořeného s pomocí procesu popsaného v Sekci 3.1. V každém časovém kroku $t - 1 \in \boldsymbol{T}$ se nacházíme ve stavu $s_{t-1} \in \boldsymbol{S}$. Po výběru akce $a_t \in \boldsymbol{A}$, popsaném v Sekci 3.3, nebo volbě náhodné akce, generujeme nový stav $s_t \in \boldsymbol{S}$ následujícím způsobem.
Vyberu z 3D pole popsaném v Sekci 3.1 diskrétní pravděpodobnostní rozdělení $p(s_t|a_t, s_{t-1})$ příslušné ke dvojici (a_t, s_{t-1}). Sestrojím distribuční funkci příslušnou k tomuto rozdělení. Následně vygenerujeme náhodnou hodnotu z intervalu $(0, 1)$ z rovnoměrného rozdělení. K této hodnotě pak najdeme vzor s užitím kvantilové funkce [5] a výslednou hodnotu přiřadím do stavu $s_t \in \boldsymbol{S}$.
Tento proces se opakuje pro všechny časové kroky $t \in \boldsymbol{T}$. Pro možnost rekonstrukce stejných výsledků při náhodném generování hodnoty z intervalu $(0, 1)$ užíváme seed, který je v každém časovém okamžiku $t \in \boldsymbol{T}$ automaticky měněn, přičítáním jedničky v každém časovém kroce $t \in \boldsymbol{T}$.

Kapitola 4

Experimentální ověření

V následující sekci představíme výsledky navrženého řešení, kde za hlavní indikátor úspěšnosti považujeme celkovou akumulovanou odměnu, neboli součet okamžitých odměn (1.48) definovaných v Sekci 2.3, za uplynulou časovou dobu $t \in \boldsymbol{T}$.
Pro verifikaci porovnáme výsledky pro tři typy strategie:

- A - optimální strategii, která používá přesný model
- B - optimální strategii, která používá odhadnutý model
- C - náhodnou strategii

Varianta A imituje plnou znalost systému, zatímco varianta B je realistický případ kde nemáme plnou znalost systému. Akce $a_t \in \boldsymbol{A}$ náhodné strategie C v čase $t \in \boldsymbol{T}$ jsou generovány dle diskrétního rovnoměrného rozdělení z množiny akcí $A \subset \mathbb{N}$. Experimenty provedeme pro strategie A, B, C .

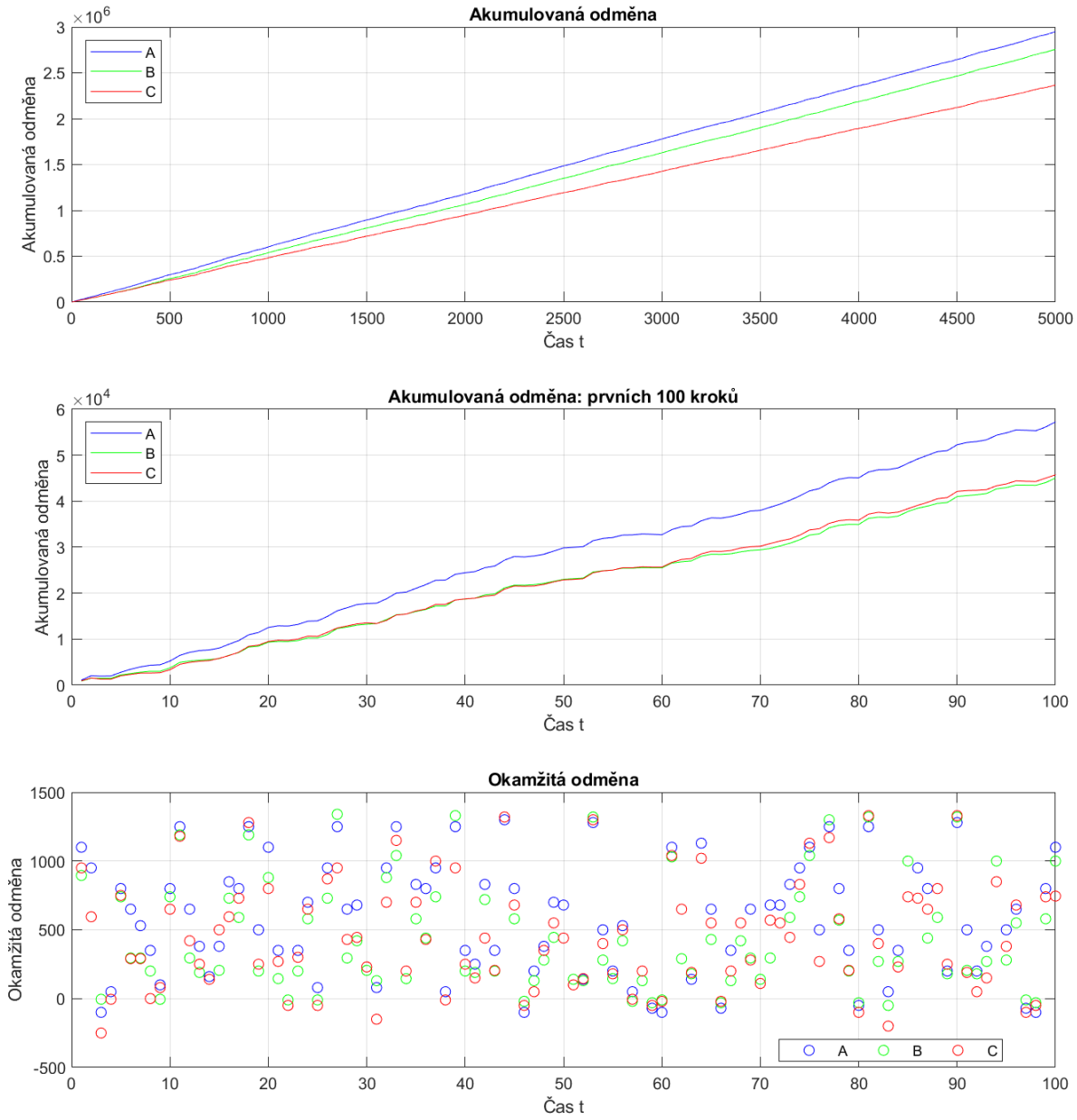
4.1 Parametry simulace

- Prostor stavů: $|\boldsymbol{S}| = 11$,
  $\boldsymbol{S} = [45000, 40000, 35000, 30000, 25000, 20000, 15000, 10000, 7000, 5000, 0]$
  (hodnoty reprezentují počty zobrazení, 0 reprezentuje prohranou aukci).
- Prostor akcí: $|\boldsymbol{A}| = 11$, $\boldsymbol{A} = [500, 400, 300, 200, 100, 70, 50, 30, 20, 10, 5]$
  (hodnoty reprezentují částku v korunách).
- Cena jednoho zobrazení = 0.03 CZK.
- počet aukcí = 5000.
- počáteční hodnota seedu = 5.

Výsledky jsou zobrazené jako časové průběhy akumulované odměny pro porovnání strategie A, B, C. Strategie A by měla být přirozeně nejlepší, jelikož operuje s přesným modelem.

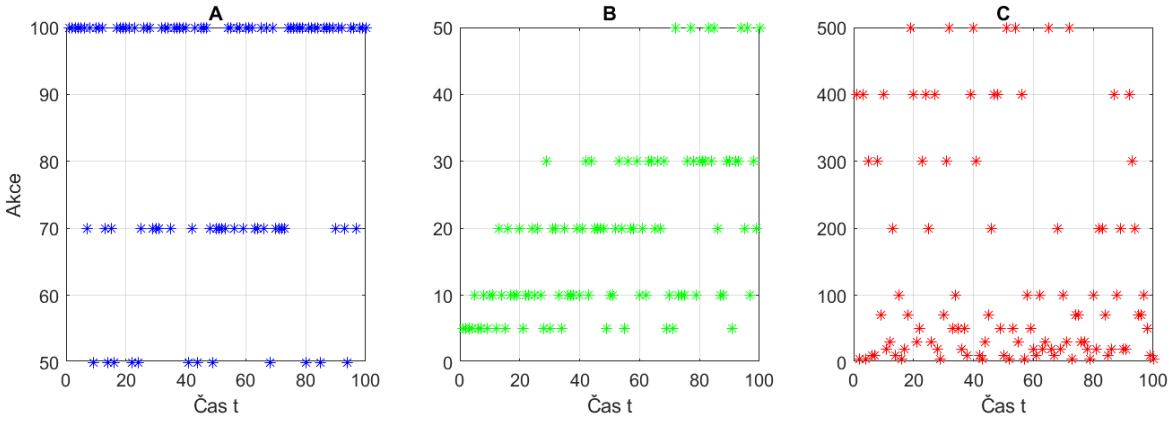
4.2 Experiment 1

Cílem Experimentu 1 je porovnat tři strategie A, B, C na 5000 časových krocích.



Obrázek 4.1: Výsledky aukce pro strategie: A, B, C

Z Obrázku 4.1 lze vidět, že při užití optimální strategie s přesným modelem (A) je výsledná akumulovaná odměna po 5000 aukcích: 2 946 850 CZK. Pří užití optimální strategie s odhadnutým modelem (B) je výsledná akumulovaná odměna 2 775 620 CZK. Což dává smysl vzhledem k tomu, že agent užívající predikovaný model má nedokonalé informace o systému a postupně jej prozkoumává. Akumulovaná odměna po 5000 aukcích pro náhodnou strategii (C) činí 2 364 855 CZK, což je méně než pro případ s užitím optimální strategie A či B.
Dále z druhého grafu na Obrázku 4.1 lze vidět, že strategie B je nejdříve horší než náhodná strategie C což je opět způsobeno nedokonalou informací o systému. S rostoucím počtem aukcí se však zlepšuje i odhadovaný model a z prvního grafu na Obrázku 4.1 lze vidět, že strategie B časem překoná strategii C.
Jak lze vidět z grafu okamžité odměny (pro ilustraci je vykresleno pouze pro prvních 100 kroků): v případě optimálních strategií A a B, dochází častěji k vysokým hodnotám okamžité odměny než pro strategii C. V případě optimální strategie A lze vidět, že dosahuje častěji vyšších hodnot okamžité odměny než optimální strategie B. To opět plyne z faktu, že v této fázi ještě není k dispozici dostatečně kvalitní odhad modelu.

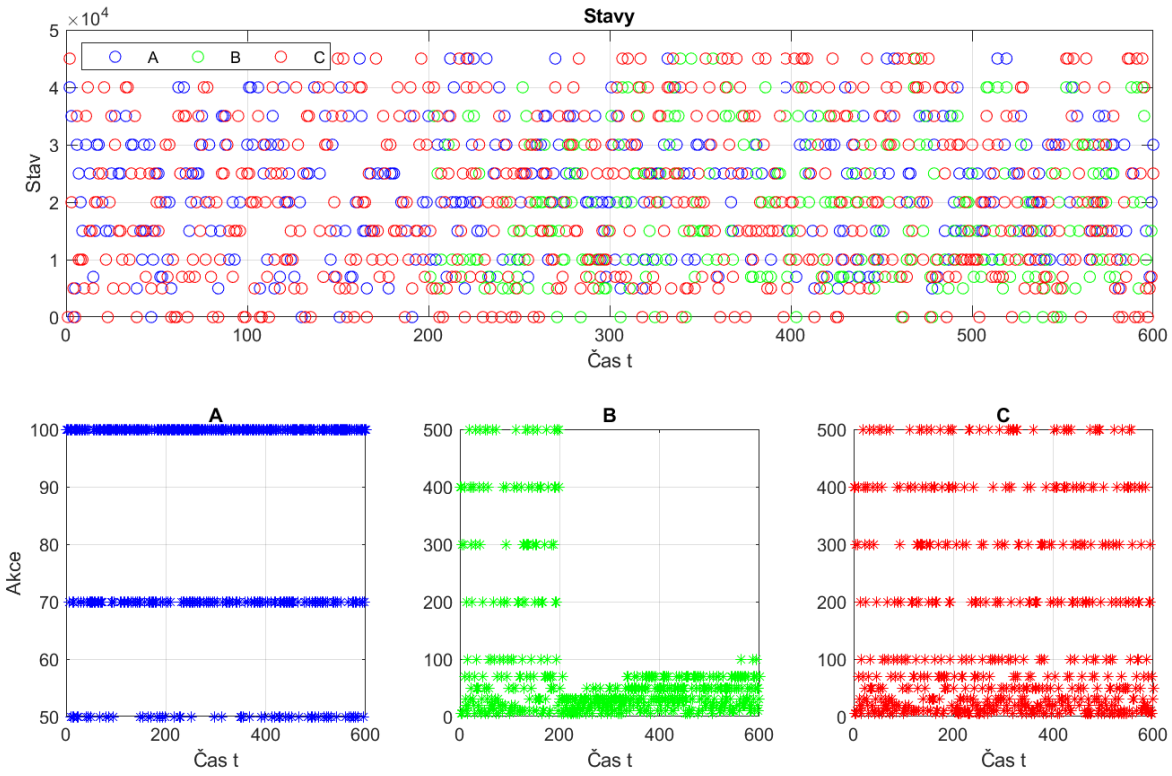


Obrázek 4.2: Časový průběh volby akcí $a_t \in \boldsymbol{A}$ pro strategie A, B, C

Z grafu na Obrázku 4.2 lze pozorovat, že při volbě optimální strategie s odhadovaným modelem (B) agent v průběhu času $t \in \boldsymbol{T}$ nejdříve volí některé akce, ale s tím jak získává lepší představu o systému tak si uvědomuje, že tyto akce nejsou výhodné. Taktéž lze vidět, že je agent opatrný, tedy že vůči volbě optimální strategie s přesným modelem (A) volí akce s nižší číselnou hodnotou.
Oproti tomu při volbě optimální strategie A lze vidět, že si je agent jistý svými volbami v průběhu času $t \in \boldsymbol{T}$ to je dáno perfektní znalostní modelu.

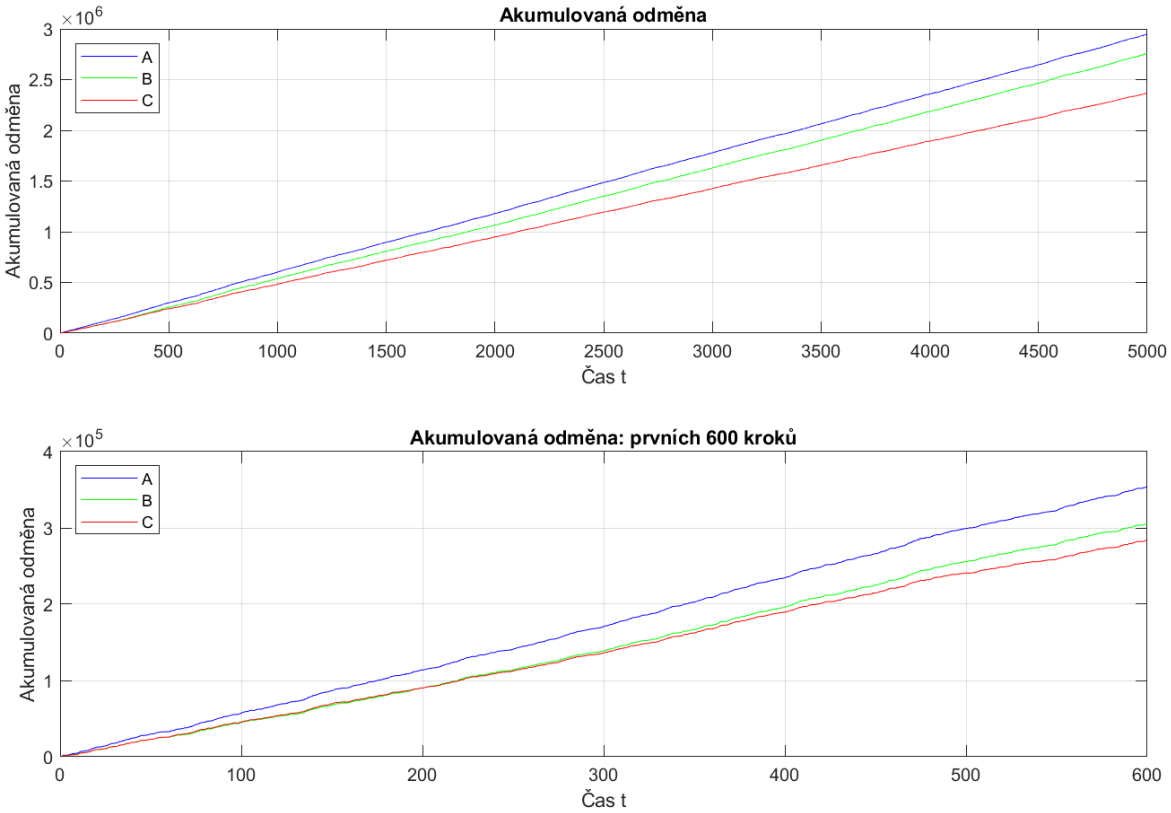
4.3 Experiment 2

V Experimentu 1 bylo využito optimální strategie s odhadovaným modelem (B), přičemž odhadování běželo v každém časovém kroce $t \in \boldsymbol{T}$. Apriorní rozdělení parametru $\theta \in \boldsymbol{\Theta}$ (1.27) bylo voleno uniformní.
Nyní nastavíme experiment stejně, jen prvních 200 časových kroků budeme užívat náhodnou strategii (C) a dále opět optimální strategii s odhadovaným modelem (B). Toto provádíme za cílem zisku lepšího aprioního odhadu pro parametr $\theta \in \boldsymbol{\Theta}$. Použití strategie C "vybudí"systém, čímž rychleji získáme informace o možných stavech. Jedná se o tzv. průzkum (exploration) v prostoru stavů. Samozřejmě za průzkum zaplatíme v krátkém běhu menší akumulovanou odměnou, ale v dlouhém běhu přinese lepší výsledky.



Obrázek 4.3: Časový průběh volby akcí $a_t \in \boldsymbol{A}$ a stavů $s_t \in \boldsymbol{S}$ pro strategie A, B, C

Z Obrázku 4.3 lze vidět, že během prvních 200 časových kroků jsou akce a stavy pro optimální strategie s odhadovaným modelem (B) a náhodnou strategii (C) identické. Po 200-stém časovém kroku se zapojí volba akce dle strategie B.

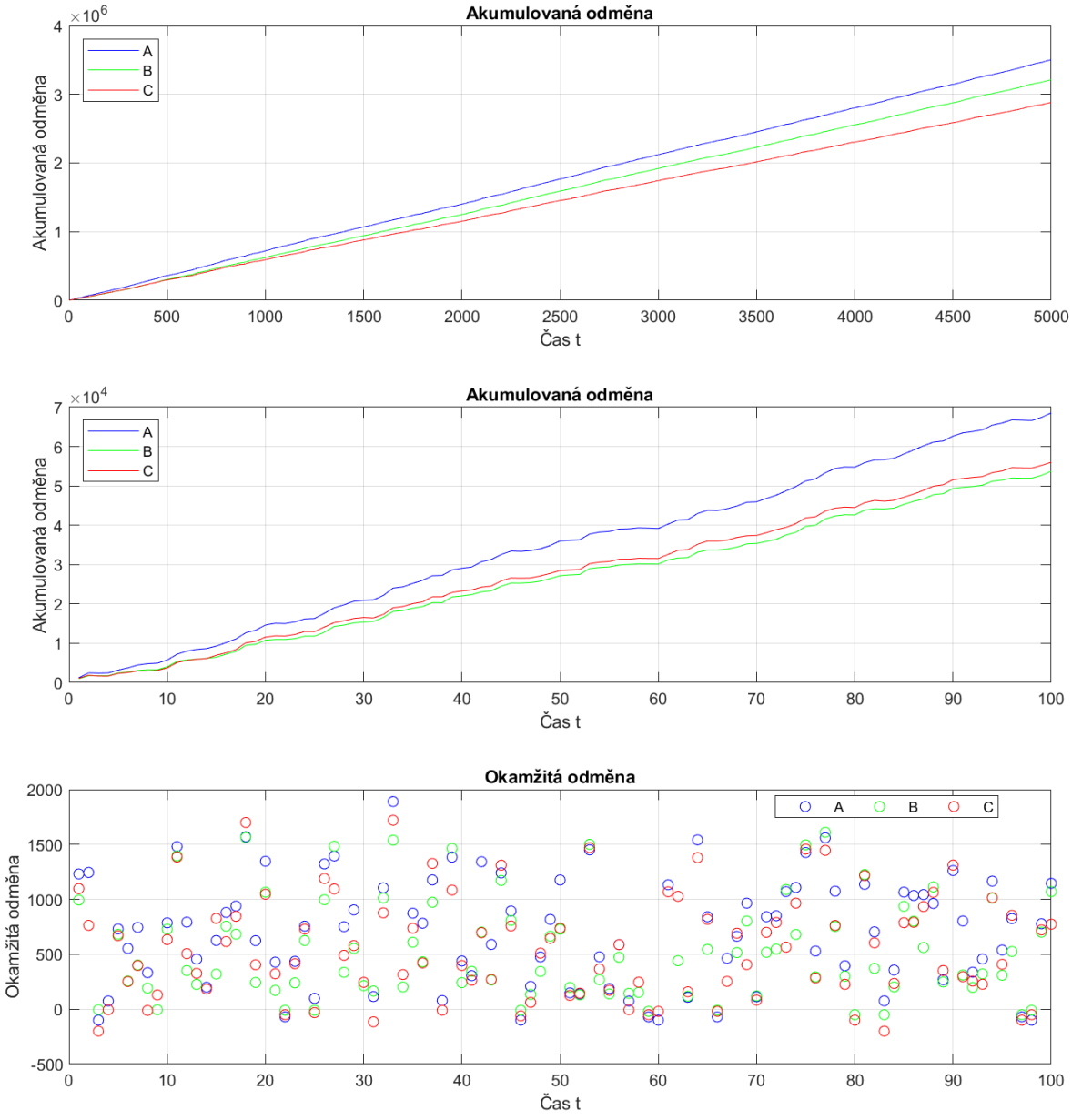


Obrázek 4.4: Akumulovaná odměna pro strategie A, B, C

Celková akumulovaná odměna pro optimální strategii s odhadovaným modelem B po 5000 aukcích činí: 2 726 625 CZK. Pro optimální strategii s generovaným modelem (A) je tato hodnota díky užití stejného seedu pro náhodnou generaci stejná jako v Experimentu 1 (2 946 850 CZK). Pro náhodnou strategii (C) je celková akumulovaná odměna po 5000 aukcích: 2 364 855 CZK, tady obě strategie A i B dosahují lepšího zisku i za podmínek tohoto experimentu.
Ve srovnání s Experimentem 1 (bez průzkumu) je hodnota celkové akumulované odměny menší než u strategie B (z Experimentu 1: 2 775 620 CZK). Toto je dáno tím, že pro stavový prostor $\boldsymbol{S} \subset \mathbb{N}_0$ a prostor akcí $\boldsymbol{A} \subset \mathbb{N}$ o velikostech $|\boldsymbol{S}| = 11$ a $|\boldsymbol{A}| = 11$ je nutné poměrně velké množství aukcí (cca 50 000) aby odhadovaný model začal reflektovat simulovaný model a tedy navržený průzkum "nebudí"systém dostatečně silně k tomu aby ztráta během průzkumu byla následně vyvážena lepší znalostí systému.

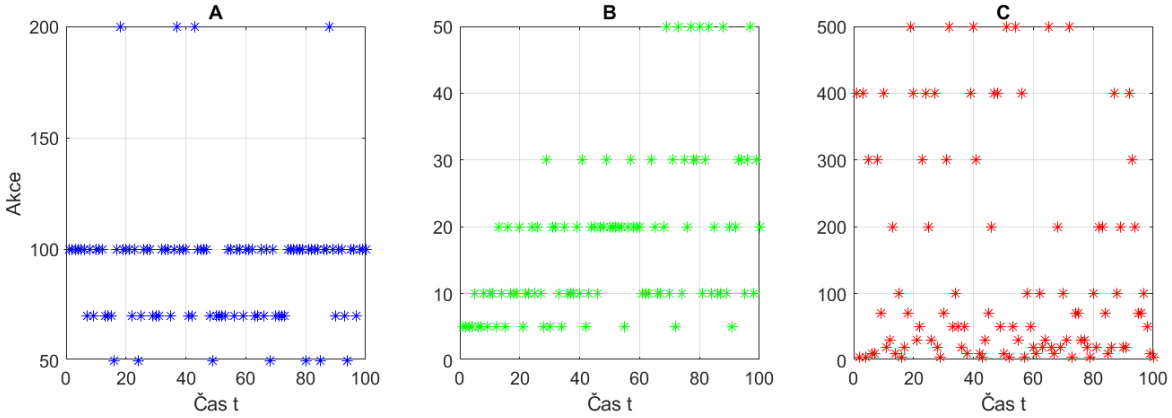
4.4 Experiment 3

Pro tento experiment zachováme parametry stejné jako v úvodu sekce, jen v každém časovém kroku $t \in \boldsymbol{T}$ budeme náhodně generovat hodnotu ceny jednoho zobrazení. Náhodné hodnoty budeme čerpat z normálního rozdělení s parametry $\mu = 0.035$ a $\sigma = 0.005$. Tato volba reflektuje následující skutečnost.
Cena jednoho zobrazení může kolísat v závislosti na pozici reklamní plochy na stránce. Vyšší cena znamená že reklamní pole je v horní části webové stránky (a při načtení stránky je zobrazená reklama přímo před očima uživatele), nižší cena reklamního pole znamená pozici v dolní části webové stránky jelikož se může stát, že se uživatel do této sekce webové stránky rozhodne nedívat. Následně pro tuto konfiguraci opět provedeme nejdříve Experiment 1 a Experiment 2.



Obrázek 4.5: Výsledky pro proměnnou cenu zhlédnuti pro strategie A, B, C

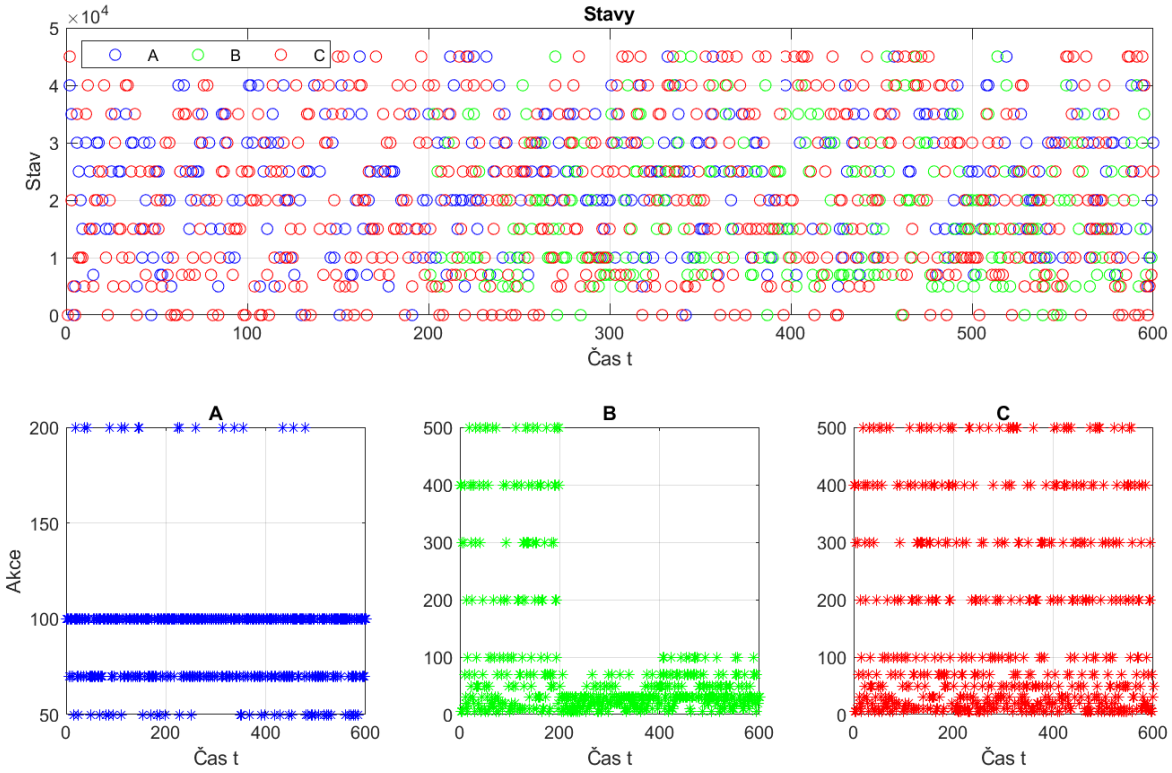
Z Obrázku 4.5 lze vidět, že při užití optimální strategie s přesným modelem (A) je výsledná akumulovaná odměna po 5000 aukcích: 3 506 242 CZK. Pří užití optimální strategie s odhadnutým modelem (B) je výsledná akumulovaná odměna 3 266 003 CZK. Což dává smysl vzhledem k tomu, že agent užívající predikovaný model má nedokonalé informace o systému a postupně jej prozkoumává. Akumulovaná odměna po 5000 aukcích pro náhodnou (C) činí 2 980 786 CZK.
Co se týče úspěšnosti algoritmu, jsou výsledky podobné jako u Experimentu 1 (bez proměnné ceny jednoho zobrazení).
Ve srovnání s experimentem 1 lze vidět, že křivka akumulované odměny více kolísá. To je dáno závislostí odměny na ceně jednoho zobrazení.
Dále z druhého grafu na Obrázku (4.1) je opět vidět, že optimální strategie B je nejdříve horší než strategie C, časem ale strategie B překoná strategii C díky učení, které v každém kroku zpřesňuje model systému.



Obrázek 4.6: Časový průběh volby akcí $a_t \in \boldsymbol{A}$ pro strategie A, B, C pro proměnnou cenu jednoho zobrazení

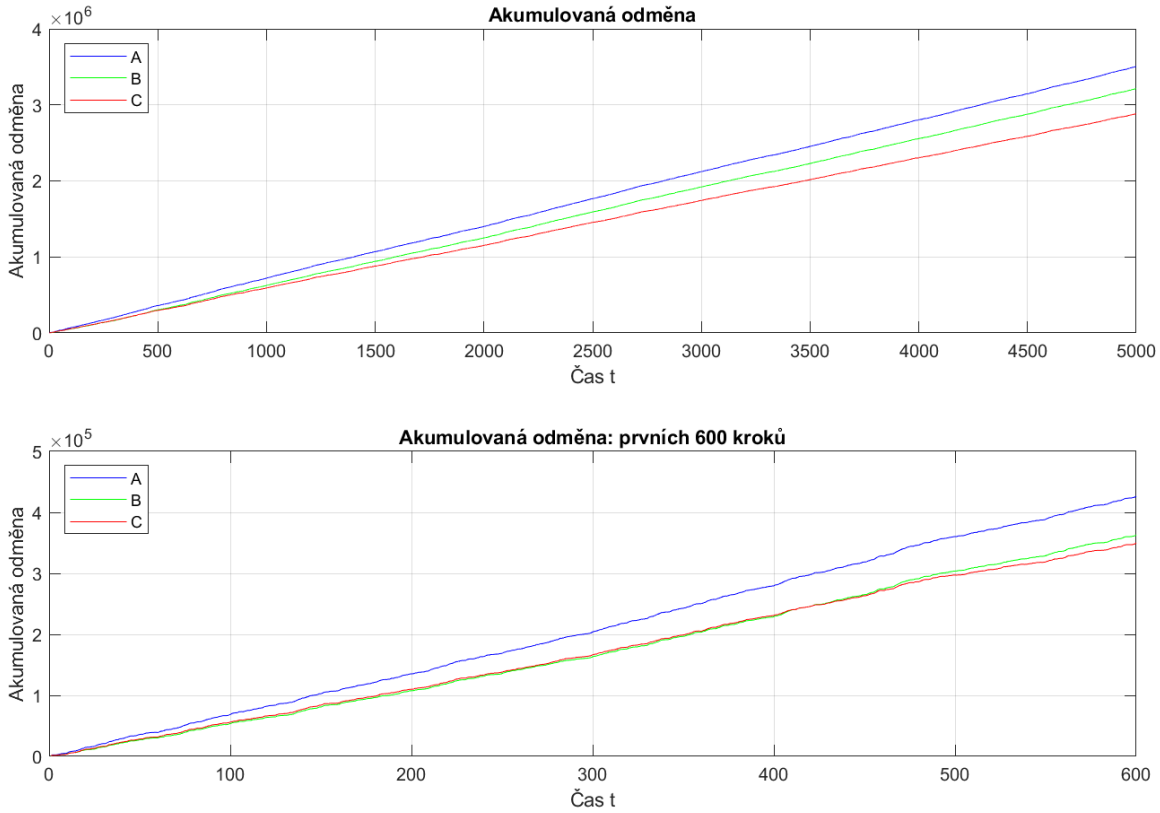
Pří srovnání obrázků 4.2 a 4.6. Lze zpozorovat, že při volbě strategie s přesným modelem (A) model má tendenci někdy volit akci 200. V Experimentu 1, s konstantní cenou jednoho zobrazení, volil maximálně 100. Agent se tedy přizpůsobil k nové situaci.

Nyní nastavíme proměnné jako v Experimentu 2, ale ponecháme proměnnou cenu jednoho zobrazení.



Obrázek 4.7: Časový průběh volby akcí $a_t \in \boldsymbol{A}$ a stavů $s_t \in \boldsymbol{S}$ pro strategie A, B, C s proměnnou cenou jednoho zobrazení

Z Obrázku 4.7 pro volbu optimální strategie s odhadovaným modelem (B) lze opět vypozorovat prvotní průzkum systému a následné zapojení strategie. Ve srovnání s obrázkem 4.3 pro volbu strategie B je agent v tomto případě opatrnější a volí akce s nižší číselnou hodnotou. To je způsobeno neurčitostí ceny jednoho zobrazení. Výsledek je podobný jako v Experimentu 2.



Obrázek 4.8: Akumulovaná odměna pro strategie A, B, C s proměnou cenou jednoho zobrazení

Celkové akumulovaná odměna pro optimální strategii s odhadovaným modelem (B) po 5000 aukcích činí 3 224 087 CZK. Pro optimální strategii s přesným modelem (A) je 3 503 469 CZK. Pro náhodnou strategii (C) je celková akumulovaná odměna po 5000 aukcích: 2 880 786 CZK, tady obě strategie A i B dosahují lepšího zisku i za podmínek experimentu 2 s proměnnou cenou jednoho zobrazení.
Výsledky jsou jinak podobné výsledkům z Experimentu 2.

Celkově je algoritmus robustní vůči kolísání ceny jednoho zobrazení i když lze očekávávat, že velké změny udělají průběh nestabilní. Tento experiment posloužil jako dobrá imitace přirozeného kolísání ceny v průběhu aukcí.

Závěr

V práci byl navržen a experimentálně ověřen algoritmus optimálního RTB s užitím odhadování. Získané výsledky byli ověřené simulačně dle navrženého modelu pro simulaci dat. Hlavním kritériem porovnání byla akumulovaná odměna. Získané výsledky ukazují, že navržená optimální strategie příhozu dává větší akumulovanou odměnu než náhodná strategie, přičemž je robustní vůči menším změnám ceny jednoho zobrazení. Učení lze o něco urychlit pomocí vhodně zvolené doby průzkumu (exploration), během kterého je akce vybírána dle náhodné strategie.
Budoucí vylepšení algoritmu zahrnují např.:

- Použití vícekrokového algoritmu volby optimální akce (v této práci byl užit pouze jednokrokový).
- Zohlednění faktu, že skutečný model je časově proměnný (např. existuje závislost na dni v týdnu, sezoně či jiných externích vlivů).
- testování na reálných datech.

Literatura

[1] S. Yuan, A. Z. Abidin, M. Sloan, and J. Wang, "Internet advertising: An interplay among advertisers, online publishers, ad exchanges and web users," 2012.

[2] S. Yuan, J. Wang, and X. Zhao, "Real-time bidding for online advertising: Measurement and analysis," 2013.

[3] Y. Cui, R. Zhang, W. Li, and J. Mao, "Bid landscape forecasting in online ad exchange marketplace," pp. 265–273, 08 2011.

[4] W. C.-H. Wu, M.-Y. Yeh, and M.-S. Chen, *Predicting Winning Price in Real Time Bidding with Censored Data*, p. 1305–1314. New York, NY, USA: Association for Computing Machinery, 2015.

[5] A. Renyi, *Probability Theory*. Dover Books on Mathematics Series, Dover Publications, Incorporated, 2012.

[6] M. L. Puterman, *Markov Decision Processes.: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[7] V. Peterka, "Bayesian approach to system identification," in *Trends and Progress in System identification*, pp. 239–304, Elsevier, 1981.

[8] J. Lin, "On the dirichlet distribution," *Department of Mathematics and Statistics, Queens University*, pp. 10–11, 2016.

[9] J. O. Berger, *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[10] D. P. Bertsekas, "Value and policy iterations in optimal control and adaptive dynamic programming," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 3, pp. 500–509, 2017.

[11] J. Wang, W. Zhang, S. Yuan, *et al.*, "Display advertising with real-time bidding (rtb) and behavioural targeting," *Foundations and Trends® in Information Retrieval*, vol. 11, no. 4-5, pp. 297–435, 2017.

[12] MATLAB, *version 9.5.0.944444 (R2018b)*. The Mathworks, Inc., Natick, Massachusetts, 2021.