Jihočeská univerzita v Českých Budějovicích
Ekonomická fakulta
Katedra aplikované matematiky a informatiky

Teze diplomové práce

# Posilované učení pro volbu trasy v rámci scénáře abstraktního provozu

**Vypracoval:** Bc. Leoš Glaser
**Vedoucí práce:** doc. Ing. Ladislav Beránek, CSc., MBA
České Budějovice
2023

---

**Klíčová slova:** posilované učení, strojové učení, umělá inteligence, doprava, návrh trasy

**Anotace:** Tato práce se zabývá přístupem posilování učení pro navrhování trasy agentovi ve zjednodušeném scénáři pohybu v dopravní síti. V teoretické části jsou představeny základy umělé inteligence, posilovaného učení a vybrané metody posilování učení. Dále je stručně zmíněna základní teorie týkající se simulace dopravy. V praktické části práce byla vytvořena konzolová aplikace využívající čtyři metody posilovaného učení. Metody byly použity pro návrh trasy svozu odpadu ve vybrané čtvrti Českých Budějovic a porovnány s metodou řešící tuto úlohu pomocí rojové inteligence. Výsledky návrhů posilovaným učením jsou podobné výsledkům získaným rojovou inteligencí, přičemž celkově nejúspěšnější metodou byla Proximal Policy Optimization s detekcí validity akcí. V jednom případu bylo nalezeno optimální řešení.

# Úvod a přehled literatury

Návrh tras je problematika, která se týká mnoha oblastí, jako je například doprava, logistika, turistika nebo robotika. Vhodně naplánované trasy mohou přispět ke snížení času na přepravu, k minimalizaci nákladů anebo ke zvýšení plynulosti dopravy. Kromě ekonomických efektů je plánovaní tras důležité mimo jiného i z hlediska životního prostředí.

Existují různé metody pro navrhování tras, od klasických algoritmů teorie grafů, přes genetické algoritmy, heuristické metody až po metody umělé inteligence. Diplomová práce se zabývá návrhem trasy pro jednoho agenta pomocí posilovaného učení. V teoretické části jsou uvedeny základy umělé inteligence, agentů a posilovaného učení včetně popisu vybraných metod, dále též základy simulací dopravy. V praktické části je navržena a vyvinuta aplikace, která je následně použita k návrhu trasy pro svoz odpadu v konkrétní části města České Budějovice.

**Umělá inteligence** má mnoho definicí a existují různé přístupy, které se mezi sebou teoreticky i prakticky více nebo méně liší. Pro téma předložené diplomové práce je významný zejména přístup z pozice *racionálního chování*, v rámci kterého hrají významnou roli racionální agenti.

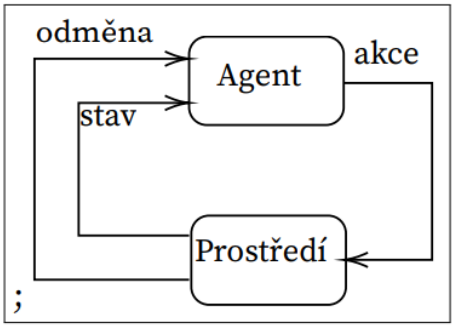
**Agentem** se rozumí cokoliv, co vnímá své prostředí prostřednictvím senzorů a působí na něj prostřednictvím akčních členů. (Norvig, 2020)

**Prostředí**, ve kterém se agent pohybuje, lze charakterizovat dle různých vlastností:

- plně pozorovatelné vs. částečně pozorovatelné vs. nepozorovatelné,
- epizodické vs. sekvenční,
- statické vs. dynamické,
- diskrétní vs. spojité,
- deterministické vs. nedeterministické,
- jednoagentní vs. multiagentní.

Při **posilovaném učení** agent vnímá prostředí, provádí akce a přijímá zpětnou vazbu ve formě *odměn*. Cílem agenta je naučit se chovat tak, aby v dlouhodobém časovém horizontu maximalizoval očekávané odměny. (Géron, 2017)

**Obrázek 1:** Základní princip posilovaného učení




***Zdroj:*** *vlastní zpracování*

Posilované učení zahrnuje dle Sutton (2020) tyto prvky: prostředí, agenta, agentovu strategii, odměnový signál, hodnotovou funkci a případně model prostředí. *Agentova strategie* (anglicky *policy*) definuje způsob chování agenta v daném čase. Mapuje stavy prostředí na akce, které mají být provedeny. *Hodnotová funkce* (anglicky *value function*) určuje, co je pro agenta žádoucí z dlouhodobého hlediska. Hodnota stavu je celková očekávaná výše odměny, kterou agent začínající v daném stavu v budoucnu získá.

K formalizaci úloh posilovaného učení slouží **Markovův rozhodovací proces** (MRP). Konečný MRP jednoho agenta sestává z diskrétní množiny stavů *S*, diskrétní množiny akcí *A*, diskrétního času *t*, pravděpodobnosti přechodu mezi stavy $p(s_i, s_{i+1}, a_i)$, odměnové funkce *R* a plánovacího horizontu *H*. (Vlassis, 2007)

**Metody posilovaného učení** lze klasifikovat dle různých charakteristik. Lapan (2018) uvádí:

- model-free vs. model-based,
- value-based vs. policy-based,
- on-policy vs. off-policy.

Model-free metody nevytváří model prostředí ani odměn, model-based metody se snaží predikovat budoucí stavy a případně i odměny. Policy-based metody přímo aproximují strategii agenta, value-based metody ohodnocují akce hodnotou a agent volí dle určitých kritérií akce. On-policy metody se snaží vyhodnocovat a zlepšovat agentovu strategii, při off-policy metodách se agent průzkumem prostředí přibližuje k optimální strategii.

Existuje mnoho metod posilovaného učení, níže jsou velmi stručně představeny vybrané z nich:

- **Q-učení**: iterativně se upravují tzv. Q-hodnoty dvojic *(stav, akce)* a algoritmus postupně konverguje k optimální strategii. (Sutton, 2020)
- **Hluboké Q-učení**: k výpočtu Q-hodnot jsou použity hluboké neuronové sítě. (Sewak, 2019)
- **SARSA:** algoritmus podobný Q-učení, avšak s odlišnou aktualizací Q-hodnot
- **Proximal Policy Optimization:** metoda upravující přímo strategii agenta pomocí úprav vah neuronové sítě. (Schulman, 2017).
- **Actor-Critic**: dvě neuronové sítě, jejichž parametry jsou upravovány – síť Actor volí akce a Critic ohodnocuje tyto volby. (Simonini, 2022)

Pro **praktickou implementaci posilovaného učení** je k dispozici řada nástrojů. Populárním je Python knihovna *Open AI Gym* poskytující rozhraní k rozličným prostředím, např. prostředí pro simulaci ATARI her, prostředí pro simulaci pohybu figur, prostředí klasických fyzikálních problémů, aj.

K vývoji aplikace v předložené diplomové práci byla použita knihovna *Stable Baselines3*, která poskytuje unifikované rozhraní pro tvorbu prostředí a nabízí řadu metod posilovaného učení.

Pro **simulaci dopravy** existuje též řada nástrojů. Obecně se dle Elefteriadou (2014) rozlišují tři typy simulací, resp. modelů dopravy:

- *Mikroskopické modely* simulují dopravu na velmi detailní úrovní aplikací vysoce detailních matematických a fyzikálních rovnic (pro pohyb vozidla, pro pohyb chodce, aj.)
- *Makroskopické modely* simulují dopravu na vyšší agregované úrovni. Doprava není simulována rovnicemi pro každé jedno vozidlo, nýbrž podobně jako tok plynu či elektřiny.
- *Mezoskopické modely* jsou kombinací makroskopických a mikroskopických modelů.

## Cíl práce

Dle zadání je cílem práce „využít přístup posilovaného učení pro návrh volbu trasy, který by se opíral pouze o zkušenosti řidičů."

Konkrétněji byl cíl praktické části stanoven na vyvinutí aplikace na bázi posilovaného učení simulující pohyb jednoho agenta v silniční síti po nejkratší okružní trase z počátečního bodu tak, aby v rámci trasy navštívil (obsloužil) předem definované body sítě. Agent na začátku nemá žádné předdefinované trasy a jednotlivé trasy volí (učí se) postupně na základě získaných zkušeností (odměn) v rámci cestování v síti. Aplikace bude otestována na úloze svozu odpadu ve vybrané lokalitě.

## Metodika

Nejdříve byla navržena aplikace. Vzhledem k použití knihovny Stable Baselines3 byl návrh omezen požadavky této knihovny. Aplikace sestává ze tří částí:

- třída `CGraph` – reprezentace silniční sítě
- třída `GraphEnv` – napsaná v souladu s požadavky na prostředí použité programové knihovny, v této třídě je definováno přidělování odměn agentovi a prováděn návrh trasy.
- učící skript – zde je definován typ agenta (resp. použité metody posilovaného učení), délka učení a uzly, které má agent obsloužit

Odměnová funkce je definována takto:

$$R(s,a) = \begin{cases} -w_n & \text{pokud } s \text{ nemá být obsluhován nebo už obsloužen byl} \\ +5 & \text{pokud } s \text{ má být obsloužen a je navštíven poprvé} \\ -5 & \text{pokud akci } a \text{ nelze ve stavu } s \text{ provést} \end{cases}$$

kde $w_n$ je normalizovaná váha hrany, přičemž hrany jsou váženy jejich délkou v metrech.

Odměnová funkce je tedy navržena tak, aby agent byl motivován navštívit všechny uzly, jež má obsloužit (obdrží v nich vysokou odměnu), zároveň aby se okolo těchto uzlů necyklil (vysokou odměnu obdrží jen při prvním navštívení) a aby tak učinil s co nejmenší ujetou vzdáleností (záporná odměna za každou další jízdu).

Algoritmus učení agenta probíhá inkrementálně v krocích. V každém kroku agent vybere akci *a* z množiny všech akcí *A* a na základě současného stavu *s* a vybrané akce obdrží odměnu. Zároveň se určí následující stav, do kterého se agent volbou akce dostane a tento stav se přidá do trasy. V případě, že trasa obsahuje všechny uzly, které měly být obslouženy, je epizoda ukončena, agent se vrátí zpět do výchozího uzlu a spočítá se výsledná nejkratší trasa epizody.

V rámci každé epizody agent postupně pohybem prostředím navštíví všechny obsluhované uzly. Tato epizodní trasa obsahuje posloupnost všech obsluhovaných uzlů. Agent však v rámci učení - zejména zpočátku, když ještě nemá mnoho naučených zkušeností - prochází síť po zbytečných redundantních smyčkách nebo zbytečně „pendluje“ mezi uzly, a proto epizodní trasa není optimální vzhledem k posloupnosti, v níž byly navštíveny obsluhované uzly. Finální navržená trasa je proto určena nejkratšími trasami mezi obsluhovanými uzly a výchozím uzlem. K určení nejkratších tras je použit Dijkstrův algoritmus.

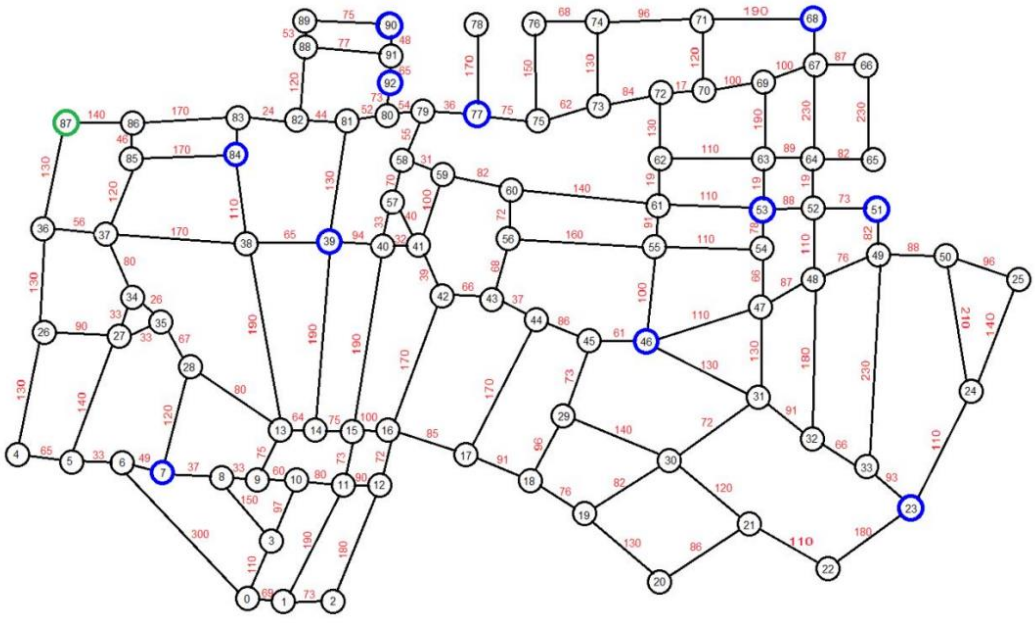
Pro učení agenta byly použity následující metody posilovaného učení:

- **A2C** – varianta Actor Critic algoritmu,
- **DQN** – metoda hlubokého Q-učení s pamětí předchozích zkušeností,
- **PPO** – algoritmus aproximující optimální strategii pomocí úprav vah neuronové sítě,
- **MPPO** – PPO s detekcí validity akcí.

Aplikace byla použita k návrhu trasy pro svozové vozidlo svážející tříděný odpad ve čtvrti Suché Vrbné v Českých Budějovicích. Tato úloha byla zvolena, protože se jedná o problém z praxe a protože byla již dříve řešena metodou rojové inteligence, s níž může být přístup posilovaného učení do určité míry porovnán.

Graf silniční sítě čtvrti Suché Vrbné je na obrázku 2, obsluhované jsou modré uzly, zelený uzel je výchozí. Červenou barvou jsou délky jednotlivých silnic.

**Obrázek 2:** Grafová reprezentace dopravní sítě čtvrti Suché Vrbné



***Zdroj:*** *převzato z* (Vácha, 2016)

Následně byly provedeny experimenty:

I. bez omezení kapacity vozidla, tj. vozidlo obslouží všechny uzly (naloží obsah všech kontejnerů) aniž by se muselo vracet do výchozího uzlu (vysypat náklad)

II. bez omezení kapacity vozidla, avšak s výpadkem jednoho spoje v polovině učení (lze připodobnit k řidiči, který musí svou „zažitou“ trasu upravit v případě uzavření silnice)

III. s omezením kapacity vozidla, po vyložení nákladu se vozidlo nevrací na původní trasu (tzn. ke zbylým dosud neobslouženým uzlům se vydá z výchozího uzlu)

IV. s omezením kapacity vozidla s návratem na původní trasu (tzn. vozidlo dojede do výchozího uzlu, vyloží náklad a vrátí se zpět do uzlu, kde bylo zaplněno)

## Výsledky

Výsledky **experimentu I.** jsou v tabulce 1 (bez případů, kdy metoda nenalezla žádné řešení). Celkově lze usoudit, že MPPO ve všech veličinách, s výjimkou směrodatné odchylky délek nalezených tras, vykazuje nejlepší výsledky a ostatní metody zcela dominuje co do počtu nalezených řešení – MPPO nalezl 99.5 % ze všech nalezených řešení celého experimentu.

**Tabulka 1**: Výsledky jednotlivých metod v případě neomezené kapacity vozidla

| Metoda | Počet kroků (v tisících) | Počet řešení | Min. trasa [m] | Max. trasa [m] | Průměr. trasa [m] | Směrod. odchylka [m] | Min. trasa – optimum [%] |
|---|---|---|---|---|---|---|---|
| **A2C** | 50 | 21 | 5 009 | 6 279 | 5 744 | 586 | 15.02 |
| | 100 | 37 | 4 941 | 6 088 | 5 877 | 593 | 13.46 |
| | 500 | 126 | 4 510 | 5 437 | 5 968 | 616 | 3.56 |
| | 1 000 | 175 | 4 709 | 5 403 | 6 065 | 662 | 8.13 |
| **DQN** | 50 | 19 | 5 134 | 6 718 | 5 991 | 491 | 17.89 |
| | 100 | 21 | 4 921 | 6 251 | 5 517 | 429 | 13.00 |
| | 500 | 23 | 4 626 | 6 895 | 5 871 | 502 | 6.22 |
| | 1 000 | 34 | 4 792 | 5 742 | 5 805 | 703 | 10.03 |
| **PPO** | 50 | 6 | 5 872 | 6 525 | 6 214 | 267 | 34.83 |
| | 100 | 25 | 4 792 | 6 613 | 5 771 | 781 | 10.03 |
| | 500 | 21 | 4 980 | 6 915 | 6 152 | 590 | 14.35 |
| | 1 000 | 20 | 5 211 | 6 939 | 6 107 | 577 | 19.66 |
| **MPPO** | 1 | 27 | 5 082 | 7 367 | 5 956 | 699 | 16.69 |
| | 10 | 104 | 4 903 | 5 570 | 6 063 | 652 | 12.58 |
| | 50 | 620 | 4 449 | 4 967 | 5 843 | 589 | 2.16 |
| | 100 | 1 791 | 4 449 | 4 825 | 5 889 | 636 | 2.16 |
| | 500 | 34 586 | 4 437 | 4 744 | 5 507 | 678 | 1.88 |
| | 1 000 | 81 605 | 4 355 | 4 722 | 5 357 | 618 | 0.00 |

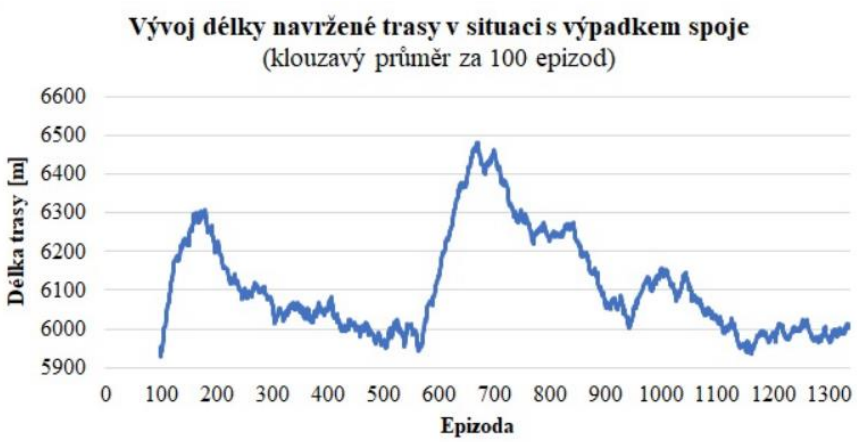
***Zdroj:*** *vlastní výsledky*

Vzhledem k dominanci metody MPPO byla v dalších experimentech používána již pouze ona.

Při **experimentu II.** byla po polovině proběhlých kroků učení „vypnuta“ hrana mezi uzly 79 a 80, protože představuje krátký spoj mezi levou a pravou částí sítě. Navíc je tato hrana blízko třem obsluhovaným uzlům a agent po odstranění hrany bude muset využít delší objížďku.

Aplikace byla spuštěna s počtem kroků 500 tisíc. Před výpadkem hrany se uskutečnilo 567 epizod s průměrnou délkou navrhnuté trasy 6 036 metrů, po výpadku 771 epizod s průměrnou délkou 6135 metrů. Na obrázku je po 567. epizodě vidět významný nárůst délky trasy, přičemž předvýpadkových délek se opět začne dosahovat až za zhruba 500 epizod.

**Obrázek 3:** Vývoj délky navržené trasy v situaci s výpadkem spoje




***Zdroj:*** *vlastní výsledky*

Pokus byl opakován s počtem kroků 1 milion. Před výpadkem se uskutečnilo 3148 epizod s průměrnou délkou navrhované trasy 5378 metrů, po výpadku 1805 epizod s průměrnou délkou trasy 6037 metrů. Narozdíl od běhu pro 500 tisíc kroků tentokrát během zbývajících epizod zůstaly délky tras až do konce běhu aplikace oproti předvýpadkovým hodnotám výrazně výše.

Výsledky **experimentu III. a IV.** jsou v tabulkách níže.

**Tabulka 2**: Výsledky v případě omezené kapacity bez návratu na původní trasu

| Metoda | Počet kroků [tis.] | Počet nalezených řešení | Min. trasa [m] | Max. trasa [m] | Průměr. trasa [m] | Směrodat. odchylka [m] |
|---|---|---|---|---|---|---|
| **MPPO** | 1 | 5 | 6 985 | 8409 | 7662 | 651 |
| | 10 | 19 | 6892 | 10270 | 8195 | 931 |
| | 50 | 15 | 7458 | 8525 | 8008 | 456 |
| | 100 | 48 | 6507 | 8457 | 8119 | 671 |
| | 500 | 54 | 7023 | 8668 | 8228 | 694 |
| | 1000 | 55 | 6310 | 8309 | 8118 | 855 |

***Zdroj:*** *vlastní výsledky*

**Tabulka 3:** Výsledky v případě omezené kapacity s návratem na původní trasu

| Metoda | Počet kroků [tis.] | Počet nalezených řešení | Min. trasa [m] | Max. trasa [m] | Průměr. trasa [m] | Směrodat. odchylka [m] |
|---|---|---|---|---|---|---|
| **MPPO** | 1 | 18 | 6029 | 7733 | 7094 | 591 |
| | 10 | 110 | 5554 | 6056 | 7017 | 714 |
| | 50 | 720 | 5538 | 6088 | 6958 | 663 |
| | 100 | 1590 | 5366 | 5715 | 6956 | 711 |
| | 500 | 29135 | 5361 | 5667 | 6521 | 779 |
| | 1000 | 56316 | 5342 | 5608 | 6776 | 593 |

**Zdroj:** vlastní výsledky

Při **srovnání s metodou rojové inteligence** (RI) je nutné poznamenat, že výsledky uvedené u metody RI byly získány pouze ze dvou běhů algoritmu, zatímco výsledky uvedené u posilovaného učení (PU) jsou pro každou jednu metodu zprůměrované z celkem 60 běhů pro 6 různých počtů kroků.

RI najde větší počet řešení než DQN a PPO, avšak je dominována metodou MPPO. Srovnatelných výsledků obě metody dosahují v průměrné délce navrhované trasy, avšak RI má menší variabilitu výsledků než PU.

**Tabulka 4:** Srovnání posilovaného učení s rojovou inteligencí

| Metoda | Varianta | Počet řešení | Min. trasa [m] | Max. trasa [m] | Průměrná trasa [m] | Směrodat. odchylka [m] |
|---|---|---|---|---|---|---|
| **Rojová inteligence** | 10 tis. iterací | 20 | 5437 | 6806 | 6065 | 384 |
| | 20 tis. iterací | 10 | 5437 | 6893 | 5976 | 490 |
| | Nejlepší | 20 | 5437 | 6806 | 5976 | 384 |
| **Posilované učení** | A2C | 13 | 5102 | 7080 | 6002 | 642 |
| | DQN | 2 | 5471 | 6114 | 5798 | 598 |
| | PPO | 1.8 | 5708 | 6311 | 6010 | 668 |
| | MPPO | 7898 | 4530 | 7782 | 5412 | 643 |

***Zdroj:*** *vlastní výsledky a výsledky převzaté z* (Vácha, 2016)

# Závěr

Předložená diplomová práce se zabývala přístupem posilovaného učení pro návrh trasy agentovi ve zjednodušeném scénáři pohybu v dopravní síti. V teoretické části byly představeny základy umělé inteligence, posilovaného učení a jeho vybraných metod. Dále byla stručně uvedena základní teorie simulování dopravy. V praktické části byla navrhnuta a vyvinuta konzolová aplikace umožňující jednoduché použití několika metod posilovaného učení.

Metody posilovaného učení byly použity na úlohu svozu odpadu a porovnány s metodou řešící tuto úlohu rojovou inteligencí. Použité metody posilovaného učení poskytly výsledky srovnatelné s rojovou inteligencí a celkově nejlepší výsledky poskytovala metoda Proximal Policy Optimization s detekcí validity akcí.

Na tuto práci by mohlo navazovat porovnání vybraných metod v různorodých scénářích provozu, dále například úprava aplikace tak, aby umožňovala i multiagentní přístup, přidání GUI anebo modifikace metod tak, aby byly škálovatelné na větší prostory stavů a akcí. Zajímavé by bylo i srovnání návrhů tras učiněných lidmi s návrhy učiněnými umělou inteligencí.

## Seznam literatury použité v tezích

Plný seznam použité literatury je uveden v textu diplomové práce.


Elefteriadou, L. (2014). *An Introduction to Traffic Flow Theory.* Springer.

Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow:*. O'Reilly Media.

Lapan, M. (2018). *Deep Reinforcement Learning Hands-On.* Packt Publishing.

Norvig, P. &. (2020). *Artificial Intelligence: A Modern Approach* (4. vyd.). Pearson.

Schulman, J. W. (2017). *Proximal policy optimization algorithms.* doi:https://doi.org/10.48550/ARXIV.1707.06347

Sewak, M. (2019). *Deep Reinforcement Learning.* Springer.

Simonini, T. (2022). *Advantage Actor Critic (A2C).* Získáno 8. 3 2023, z https://huggingface.co/blog/deep-rl-a2c

Sutton, R. &. (2020). *Reinforcement Learning: An Introduction* (2. vyd.). The MIT Press.

Vácha, L. (2016). *Řešení optimální cesty svozu odpadů pomocí rojové inteligence.* diplomová práce, Jihočeská univerzita v Českých Budějovicích.

Vlassis, N. (2007). *Concise Introduction to Multiagent Systems and Distributed Artificial.* Morgan & Claypool Publishers. Načteno z https://jmvidal.cse.sc.edu/library/vlassis07a.pdf

# Další bibliografické údaje

| | |
|---|---|
| **Type of thesis:** | master |
| **Author:** | Leoš GLASER |
| **Academic Year:** | 2022/2023 |
| **Department:** | Department of Applied Mathematics and Informatics, Faculty of Economics, University of South Bohemia in Ceske Budejovice, Czech Republic |
| **Thesis supervisor:** | doc. Ing. Ladislav Beránek, CSc., MBA |
| **Number of pages:** | 75 |
| **Language od thesis:** | CZ |
| **Title of Thesis:** | Posilované učení pro volbu trasy v rámci scénáře abstraktního provozu |
| **Annotation:** | This thesis deals with reinforcement learning approach for designing a route for an agent in a simplified scenario of movement in a transportation network. The theoretical part introduces the basics of artificial intelligence, reinforcement learning, and selected methods of reinforcement learning, both classical and modern. Additionally, the basic theory related to traffic simulation is briefly mentioned. In the practical part of the thesis, a console application utilizing four reinforcement learning methods was developed. These methods were used to design a waste collection route in a selected district in České Budějovice and compared to a method solving this task using swarm intelligence. The results of the reinforcement learning-designed routes are similar to the results obtained by swarm intelligence, with Proximal Policy Optimization with action validity detection being the most successful method overall. In one case, an optimal solution was found. |
| **Key words:** | reinforcement learning, machine learning, artificial intelligence, traffic, route design |

---

Přílohy volně vložené: žádné

Přílohy vázané v práci: žádné