ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	**Limitations of Reinforcement Learning Algorithms in Imperfect Information Games**
Jméno autora:	**Bc. Jakub Koubele**
Typ práce:	diplomová
Fakulta/ústav:	Fakulta elektrotechnická (FEL)
Katedra/ústav:	Katedra počítačů
Oponent práce:	Mgr. Tomáš Gavenčiak, PhD
Pracoviště oponenta práce:	Independent researcher

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání — náročnější

Hodnocení náročnosti zadání závěrečné práce.

Zadání je spíše náročnější, směřující k otevřenému vědeckému problému, nastiňuje konkrétní rámec práce alvšak nechává dost prostoru pro volbu konkrétních problémů i řešení.

Splnění zadání — splněno

Posuďte, zda předložená závěrečná práce splňuje zadání. V komentáři případně uveďte body zadání, které nebyly zcela splněny, nebo zda je práce oproti zadání rozšířena. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.

Body zadání jsou splněny dostatečně. Review literatury o RL ve hrách je přítomen jen v základní formě, avšak práce obsahuje teoretické rozpracování, experimentální evaluaci i nové teoretické výsledky, spolu s dobrým úvodem do užitých technik a konceptů.

Zvolený postup řešení — správný

Posuďte, zda student zvolil správný postup nebo metody řešení.

Viz též posudek níže.

Odborná úroveň — A - výborně

Posuďte úroveň odbornosti závěrečné práce, využití znalostí získaných studiem a z odborné literatury, využití podkladů a dat získaných z praxe.

Velmi dobrá, viz též posudek níže.

Formální a jazyková úroveň, rozsah práce — A - výborně

Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku.

Úroveň zpracování je velmi dobrá po odborné, jazykové i prezentační stránce. Chyby v gramatice a typografii se vyskytují, ale poměrně zřídka a nijak nevadí obsahu.

Výběr zdrojů, korektnost citací — B - velmi dob

Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení závěrečné práce. Charakterizujte výběr pramenů. Posuďte, zda student využil všechny relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.

Práce využívá citace právně, avšak na svůj rozsah ve spíše menším množství a oproti zadání neobsahuje rešerši oblasti.

Další komentáře a hodnocení
Vyjádřete se k úrovni dosažených hlavních výsledků závěrečné práce, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, publikačním výstupům, experimentální zručnosti apod.
Viz posudek níže.

III. CELKOVÉ HODNOCENÍ, OTÁZKY K OBHAJOBĚ, NÁVRH KLASIFIKACE

Shrňte aspekty závěrečné práce, které nejvíce ovlivnily Vaše celkové hodnocení. Uveďte případné otázky, které by měl student zodpovědět při obhajobě závěrečné práce před komisí.

Viz posudek níže.

Předloženou závěrečnou práci hodnotím klasifikačním stupněm A - výborně.

Datum: 25. 1. 2020

Podpis: Mgr. Tomáš Gavenčiak, PhD

Opponent review for

Limitations of Reinforcement Learning Algorithms in Imperfect Information Games

The thesis examines the performance of common reinforcement learning methods on imperfect information games. In these scenarios, RL algorithms provide no performance guarantees and have been observed to fail, yet are sometimes used in practice. Further examination of the shortcomings of RL algorithms in general games and building theory behind them is therefore an important goal, and this thesis makes good progress in this direction.

The thesis is very well written and structured, gives a good introduction to the area, concepts and techniques, clearly presents relevant existing and novel theoretical results, and presents experimental results. The author finds novel results (e.g. Thm 2.2, 3.1 and 4.2) and presents theoretical derivation of their algorithms (e.g. Algorithm 7). The mathematics and theory are on a very high level, although some formalities are not always clear. I enjoyed reading the thesis and while I could find several technical comments and questions, I evaluate the thesis very favorably and encourage the author in pursuing this line of research and continuing their scientific education.

Comments

These comments are not meant as an extensive list and are meant as prompts for discussion rather than outright criticism.

The experiments from chapter 5 conclude that, against a fixed opponent, Q-learning does not learn a close-to-optimal strategy and the plotted learning progress seems like a convincing evidence.

For self-play, neither Q-learning nor Policy Gradient learn a close-to-optimal strategy, however, I think the evidence of non-convergence would be more convincing. My main objection is that the quality of time-averaged strategies are generally hyperbolic, e.g. $1/T^{\alpha}$, and the time frame of the experiments may be too short for smaller learning rates (e.g. $\eta<0.05$ in the case of Policy Gradient, $\eta<0.01$ obviously has not converged).

Section 4.2, while valuable, could be improved with more structure, e.g. stating some parts as lemmas, stating the goal of the derivation etc. Here and elsewhere, it could be beter mentioned what are novel contributions and what is recapitulation or re-derivation of known properties.

I find the result of Thm 2.2 very neat. I would not be too surprised if this was already known but can‘t verify this myself. I find an independent derivation of the theorem interesting in any case.

In Section 4.2 before (50): There are more necessary conditions for a given function to be a gradient of some F, namely the converse of the Gradient theorem. Not sure whether the existence of F is crucial for the alorithm in the thesis. May be also relevant in proof of Thm 4.2.

The result tables presented in chapters 3 and 4 are hard to interpret, although some observations are provided by the author. E.g. Table 6 would be neatly presented by a graph of Max. diff. Depending on $\eta T^{0.5}$, similarly for other tables.

I would expect the smaller values of η to perform strictly better than large η if given time to converge (unless some fiddly functional approximator, e.g. a neural net, is used). This seems to be supported e.g. by Table 7, games 10x10, where the performance of large η stagnates while the performances of smaller η improve. E.g. for $T=10^7$, the optimal η moves toward the smaller ones. In this light, exploring many values of learning rate in Ch 4 may not be as useful, and more exposition for the extreme values could be better.

Questions for the author

- Theorem 2.2 talks only about zero-sum games. What are the obstacles to formulating a similar theorem for general (non-zero-sum) games? E.g. are there examples where the two metrics differ arbitrarily?
- (optional) In Fig 5, it seems that larger learning rates have converged to various Nash distances, while the smaller ones may have not converged. If computationally feasible, how would the figure look for a much larger number of steps, e.g. 10^7 or 10^8?
- Do you plan to refine this work into a publication?

Mgr. Tomáš Gavenčiak, PhD
25 January 2020 in Brno