Prof. dr hab. inż. Leszek Rutkowski

Instytut Badań Systemowych PAN

Warszawa, 30.03.2026

RECENZJA ROZPRAWY DOKTORSKIEJ

1. Informacje ogólne

Tytuł rozprawy: Sample-Efficient Actor-Critic Algorithms in Reinforcement Learning (Wydajność próbkowa algorytmów typu aktor-krytyk w uczeniu ze wzmocnieniem)

Autor: mgr Michał Filip Nauman

Promotor: dr hab. Marek Cygan, prof. Uniwersytetu Warszawskiego

Jednostka prowadząca: Rada Naukowa Dyscypliny Informatyka, Uniwersytet Warszawski

Dyscyplina: Dziedzina nauk ścisłych i przyrodniczych, dyscyplina: informatyka

Zgodnie z wymogami ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce, art. 187, recenzja obejmuje: (1) ocenę ogólnej wiedzy teoretycznej doktoranta, (2) ocenę umiejętności samodzielnego prowadzenia pracy naukowej oraz (3) ocenę oryginalności rozwiązania naukowego.

W części pierwszej, rozprawa doktorska mgr. inż. Michała Naumana pt. Sample-Efficient Actor-Critic Algorithms in Reinforcement Learning, przygotowana pod kierunkiem dr. hab. Marka Cygana, prof. Uniwersytetu Warszawskiego, łączy problematykę głębokiego uczenia ze wzmocnieniem, efektywności próbkowania i skalowalności algorytmów aktor-krytyk. Bez wątpienia jest to jeden z najbardziej dynamicznie rozwijających się nurtów współczesnej informatyki.

Rozprawa obejmuje 35 stron maszynopisu zawierających: streszczenia w językach angielskim i polskim, wprowadzenie do uczenia ze wzmocnieniem i algorytmów aktor-krytyk, motywację badań, omówienie wkładu naukowego, dokładny opis zawartości załączonych sześciu prac (P1-P6), wnioski i kierunki przyszłych badań, podziękowania oraz spis 87 pozycji literatury. W drugiej części zamieszczono odbitki sześciu artykułów opublikowanych na konferencjach: ICML 2023, ICML 2024, AAAI 2025, IJCAI 2025, NeurIPS 2024 (Spotlight, top 3%) oraz ICML 2025.

Wykaz prac stanowiących zasadniczą część rozprawy:

[P1] M. Nauman, M. Cygan. On many-actions policy gradient. ICML 2023.

[P2] M. Nauman, M. Bortkiewicz, P. Miłos, T. Trzcinski, M. Ostaszewski, M. Cygan. Overestimation, overfitting, and plasticity in actor-critic: the bitter lesson of reinforcement learning. ICML 2024.

[P3] M. Nauman, M. Cygan. Decoupled policy actor-critic: bridging pessimism and risk awareness in reinforcement learning. AAAI 2025.

[P4] M. Nauman, M. Ostaszewski, M. Cygan. A case for validation buffer in pessimistic actor-critic. IJCAI 2025.

[P5] M. Nauman, M. Ostaszewski, K. Jankowski, P. Milos, M. Cygan. Bigger, regularized, optimistic: scaling for compute and sample-efficient continuous control. NeurIPS 2024 (Spotlight, top 3%).

[P6] O. Rybkin, M. Nauman, P. Fu, C. Snell, P. Abbeel, S. Levine, A. Kumar. Value-based deep RL scales predictably. ICML 2025.

2. Motywacja i ogólna idea rozprawy

Głównym celem pracy jest zwiększenie wydajności próbkowej (ang. *sample efficiency*) algorytmów typu aktor-krytyk w głębokim uczeniu ze wzmocnieniem (RL – *Reinforcement Learning*). Wydajność próbkowa określa , jak wiele interakcji ze środowiskiem potrzebuje agent, aby nauczyć się optymalnego, w określonym sensie, zachowania. Jest to kluczowe w zastosowaniach rzeczywistych (np. robotyce), gdzie pozyskanie każdej dodatkowej próbki może być kosztowne. Autor argumentuje, że choć RL odniosło sukcesy (np. gry wideo, sterowanie robotami), wciąż boryka się z niestabilnością i ogromnym zapotrzebowaniem na dane. Rozprawa łączy teoretyczne analizy z praktycznymi algorytmami, kładąc szczególny nacisk na to, że metody znane z innych dziedzin uczenia maszynowego (skalowanie modeli, regularyzacja) są często skuteczniejsze niż ręcznie projektowane usprawnienia dedykowane dla RL. W szczególności rozprawa jest osadzona w rodzinie algorytmów aktor-krytyk: „aktor" to część, która wybiera akcje (polityka), a „krytyk" to część, która ocenia, jak dobre są stany i akcje . Jeśli krytyk ocenia niestabilnie albo myli się systematycznie (np. zawyża oceny), agent podejmuje złe decyzje i potrzebuje dużo prób, żeby się „oduczyć" błędów. Jeśli aktor ma gradient o dużej wariancji (szum w sygnale uczenia), to również potrzeba dużo danych, by zorientować się, w którą stronę aktualizować strategię. Cała praca jest więc zbiorem spójnych odpowiedzi na pytanie: co konkretnie zrobić w konfiguracji w aktor-krytyk, żeby szybciej i pewniej uczyć się na podstawie małej liczby próbek. W załączonych publikacjach jest to realizowane z wykorzystaniem teorii wariancji gradientu polityki, analizę stabilności i regularyzację w metodach *off-policy*, odpowiednie balansowanie optymizmu i pesymizmu, monitorowanie błędu aproksymacji krytyka oraz skalowanie modeli i budżetu obliczeniowego w określony sposób. Opiniowana rozprawa doktorska mieści się w nurcie uczenia maszynowego "*Data Centric* ML". **Jest to jedna z najlepszych rozpraw doktorskich, jakie dotychczas opiniowałem.**

3. Ocena ogólnej wiedzy teoretycznej doktoranta

Bez wątpienia mgr. inż. Michał Nauman dowodzi rozległej i głębokiej wiedzy teoretycznej z zakresu informatyki, a w szczególności głębokiego uczenia maszynowego i uczenia ze wzmocnieniem. Doktorant wykazuje biegłą znajomość formalnego aparatu procesów decyzyjnych Markowa (MDP), teorii gradientów polityki, algorytmów różniczkowania

temporalnego (TD), a także architektury i treningu głębokich sieci neuronowych. Należy odnotować znajomość przez doktoranta zaawansowanej matematyki stosowanej, jak np. twierdzenia o centralnym twierdzeniu granicznym dla łańcuchów Markowa, analizy wariancji estymatorów stochastycznych gradientów polityki, teorii użyteczności (Von Neumann-Morgenstern), a także teorii programowania dynamicznego w kontekście analizy błędów aproksymacji krytyka. Wiedza ta jest twórczo zastosowana do wyprowadzenia oryginalnych wyników teoretycznych , w szczególności warunków optymalności próbkowania wielu akcji (Twierdzenie 3.3 w P1) oraz analizy błędu pesymistycznego krytyka (P4). Doktorant wykazuje znajomość szerokiego spektrum literatury: 87 pozycji bibliograficznych obejmuje kluczowe prace z uczenia ze wzmocnieniem (Sutton i Barto, Mnih i in., Silver i in., Haarnoja i in., Fujimoto i in.), teorii skalowania modeli (Kaplan i in.), a także teorii decyzji. W wielu pracach doktorant nawiązuje do przełomowych rezultatów z innych dziedzin (NLP, computer vision) i twórczo przenosi ich metodologię do RL, co świadczy o interdyscyplinarności i szerokości horyzontów naukowych kandydata. Niezależnie od silnego fundamentu teoretycznego, doktorant wykazuje równie imponującą wiedzę empiryczną i inżynierską: projektowanie skalowalnych eksperymentów (ponad 60 wariantów agentów w P2, dziesiątki tysięcy godzin obliczeń na klastrach GPU), tworzenie repozytoriów kodu, a także umiejętność krytycznej analizy wyników statystycznych (zastosowanie miar IQM, przedziały ufności bootstrapowe zgodne z metodyką RLiable). Podsumowując, recenzowana rozprawa prezentuje wiedzę teoretyczną i praktyczną wyraźnie przekraczającą standardowy poziom oczekiwany od doktora nauk - zarówno w zakresie formalnego aparatu matematycznego, jak i znajomości stanu wiedzy w zakresie uczenia maszynowego.

4. **Ocena umiejętności samodzielnego prowadzenia pracy naukowej**

Rozprawa doktorska jednoznacznie i przekonująco dowodzi umiejętności samodzielnego prowadzenia badań naukowych przez mgr. Naumana. Spośród sześciu prac wchodzących w skład dysertacji, doktorant szacuje swój wkład na 85% w pracach P1 i P3 (napisanych we współautorstwie wyłącznie z promotorem), 75% w P5, 60% w P4 oraz 40% w P2. Jedynie w P6 rola kandydata była bardziej ograniczona (15%). Jednak ta praca powstała we współpracy z renomowanym ośrodkiem (UC Berkeley) i stanowi mocny akcent rozprawy. Doktorant wykazuje wszystkie cechy dojrzałego naukowca: identyfikuje istotne i otwarte problemy naukowe, formułuje precyzyjne hipotezy badawcze, stosuje adekwatną metodologię, wyciąga wnioski poparte zarówno teorią, jak i eksperymentami i potrafi je przedstawiać w formie przyjętej przez prestiżowe konferencje. Wartym podkreślenia jest fakt, że praca P5 została wyróżniona Spotlightem na NeurIPS 2024 (top 3% przyjętych prac), co jest wyjątkowym osiągnięciem dla doktoranta. Doktorant wykazuje samodzielność nie tylko w realizacji eksperymentów, ale również w zakresie tworzenia nowych koncepcji algorytmicznych (MBMA, DAC, VPL, BRO) i ich formalnego uzasadnienia. Połączenie teorii decyzji z algorytmami aktor-krytyk (P3) jest przykładem oryginalnego rozwiązania, które doktorant przedstawił we współpracy z promotorem. W trakcie realizacji rozprawy doktorant odbył

dziewięciomiesięczny staż w grupie prof. Pieter Abbeel na Uniwersytecie Kalifornijskim w Berkeley, gdzie nawiązał współpracę z czołowymi badaczami (Sergey Levine, Aviral Kumar), co zaowocowało publikacją P6 na ICML 2025. Świadczy to o zdolności do funkcjonowania w międzynarodowym środowisku naukowym

5. **Ocena oryginalności rozwiązania naukowego.**

Poniżej krótko przedstawiam motywacje dla podjęcia tematyki w każdej z 6 prac oraz główne rezultaty.

[P1] On Many-Actions Policy Gradient

W standardowym uczeniu ze wzmocnieniem agent, będąc w danym stanie, wybiera jedną akcję, obserwuje wynik i uczy się. Autorzy zadają pytanie: Jak wyglądałby proces decyzyjny , gdyby agent mógł w danej chwili rozważać wiele możliwych akcji naraz, obliczać gradient uczenia po wszystkich z nich i dopiero wtedy wykonać rzeczywisty krok? Praca dowodzi matematycznie, że dla wielu rodzajów problemów decyzyjnych (procesów Markowa) takie wieloakcyjne próbkowanie zmniejsza wariancję gradientu polityki i prowadzi do szybszej nauki. Autorzy proponują algorytm MBMA (Model-Based Many-Actions), który korzysta z nauczonych modeli dynamiki środowiska do symulowania tych dodatkowych akcji, bez konieczności faktycznego ich wykonywania. Eksperymenty potwierdzają skuteczność podejścia.

[P2] Overestimation, Overfitting, and Plasticity — Bitter Lesson of RL

Algorytmy RL borykają się z wieloma poważnymi problemami, m.in. (a) przeszacowaniem nagród , agent błędnie ocenia, że dane akcje są lepsze niż są w rzeczywistości; (b) przeuczeniem (overfitting), sieć neuronowa "zapamiętuje" przeszłe sytuacje zamiast uogólniać; (c) utratą płastyczności, po długim treningu sieć staje się "sztywna" i trudno ją dalej optymalizować, Autorzy przeprowadzają systematyczne badanie 9 technik radzenia sobie z tymi problemami w ramach algorytmu Soft Actor-Critic (SAC), testując ponad 60 konfiguracji. Kluczowe odkrycie: ogólne techniki regularyzacji sieci neuronowych (layer normalization, spectral normalization) są skuteczniejsze od specjalistycznych metod uczenia ze wzmocnieniem.

[P3] Decoupled Policy Actor-Critic (DAC)

W standardowych algorytmach RL stosuje się pesymistyczne szacowanie wartości (agent zakłada, że sytuacja jest gorsza niż być może jest), by uniknąć przeoszacowania. Problem w tym, że zbyt pesymistyczny agent nie eksploruje wystarczająco. Autorzy po raz pierwszy w literaturze łączą pesymizm i optymizm w uczeniu ze wzmocnieniem z teorią oczekiwanej użyteczności zaczerpniętej z ekonomii behawioralnej. Proponują algorytm DAC z dwoma aktorami: pesymistycznym (ostrożnie działającym) i optymistycznym (eksplorującym). Pesymistyczny aktor uczy krytyka, optymistyczny zbiera dane.

[P4] Validation Buffer in Pessimistic Actor-Critic (VPL)

Autorzy proponują rozwiązanie analogiczne do walidacji krzyżowej w uczeniu maszynowym: część danych jest odłożona do osobnego "buforu walidacyjnego", służącego wyłącznie do oceny błędu aproksymacji krytyka. Praca dowodzi, że błąd aproksymacji krytyka uczonego metodą różnicy temporalnej można modelować rekurencyjnie, podobnie jak równanie Bellmana. Na tej podstawie opracowano algorytm VPL, który dynamicznie dostosowuje poziom pesymizmu, minimalizując błąd aproksymacji bez przeuczenia.

[P5] BRO - Bigger, Regularized, Optimistic (NeurIPS 2024 Spotlight)

W przetwarzaniu języka naturalnego i wizji komputerowej okazało się, że większe modele prowadzą do lepszych (w określonym sensie) wyników. Ale w uczeniu ze wzmocnieniem naiwne powiększanie sieci neuronowych prowadzi do niestabilności i gorszych wyników. Problem polega na tym , że RL uczy się na niestacjonarnych danych, przez co duże modele łatwo się "roztrajają". Autorzy proponują architekturę BroNet wraz z algorytmem BRO, który łączy: większe modele (skalowanie parametrów), regularyzacje (dla stabilności) i optymistyczną eksplorację. Eksperymenty na 40 złożonych zadaniach wykazują, że BRO osiąga wyniki lepsze niż najlepszy wówczas algorytm znany w literaturze, korzystając przy tym ze znacznie zmniejszonego czasu obliczeniowego. Praca uzyskała wyróżnienie Spotlight na NeurIPS 2024 (top 3% zgłoszeń).

[P6] Value-Based Deep RL Scales Predictably (ICML 2025)

W uczeniu nadzorowanym odkryto tzw. Prawa skalowania (scaling laws): podwojenie danych lub parametrów modelu poprawia wyniki w przewidywalny, matematyczny sposób (prawo potęgowe). Autorzy zadają pytanie: czy podobne prawa obowiązują w uczeniu ze wzmocnieniem? To fundamentalne pytanie, bo pozwoliliby planować zasoby obliczeniowe, tak jak w dużych modelach językowych. Autorzy formalizują problem skalowania RL jako zagadnienie optymalizacji wielu zasobów: danych i obliczeń. Wprowadzono pojęcie „replay ratio" i pokazano, że dane i obliczenia tworzą granicę Pareto, co umożliwia wymianę danych na obliczenia i odwrotnie w przewidywalny sposób, zgodnie z prawem potęgowym. Pozwala to efektywnie dobierać hiperparametry sieci (batch size, learning rate) bez kosztownych przeszukiwań siatki.

6. Uwagi o charakterze dyskusyjnym

Rozprawa podejmuje fundamentalne i aktualne wyzwanie informatyki: jak sprawić, by algorytmy uczenia ze wzmocnieniem uczyły się szybciej, używając mniej prób z otoczeniem. Wszystkie sześć prac wnosi oryginalny wkład naukowy w rozwój dyscypliny naukowej informatyki. Poniższe uwagi mają charakter dyskusyjny i w niczym nie umniejszają tego wkładu.

a) Algorytm MBMA wymaga nauczenia dodatkowego modelu dynamiki środowiska, co wprowadza dodatkowy koszt obliczeniowy i ryzyko błędu modelowania. W przypadku środowisk z wysokowymiarowymi przestrzeniami stanowisk nauczenie dokładnego modelu dynamiki może być bardzo trudne. Eksperymenty prowadzono głównie na standardowych

zadaniach testowych znanych w literaturze. Weryfikacja praktyczna (np. robotyka rzeczywista) byłaby cenna. Analiza teoretyczna opiera się na ergodyczności MDP, co może ograniczać stosowalność w rzeczywistych problemach.

b) Praca P2 jest pierwszą w literaturze tak systematyczną analizą synergii między technikami redukcji przeoczenia, overfittingu i plastyczności w ramach jednego ujednoliconego frameworku. Badanie jest czysto empiryczne, bez komplementarnych wyników teoretycznych wyjaśniających mechanizm obserwowanych synergii. Testy prowadzono na środowiskach , które mogą nie w pełni odzwierciedlać złożoności rzeczywistych zastosowań. Liczba badanych technik stanowi wybraną podprzestrzeń możliwości.

c) Zastosowanie dwóch aktorów w pracy P3 zwiększa złożoność pamięciową i obliczeniową algorytmu. Analiza teoretyczna opiera się na upraszczających założeniach, które mogą nie być spełnione w bardziej skomplikowanych problemach. Hiperparametry pesymizmu i optymizmu wymagają strojenia, co może być trudne w praktyce.

d) Podział danych w pracy P4 na bufor treningowy i walidacyjny efektywnie zmniejsza ilość danych dostępnych dla każdego z celów. W środowiskach z ograniczoną liczbą próbek może to być kosztowne. Wybór proporcji podziału buforu wymaga strojenia i jest wrażliwy na charakterystykę środowiska. Wyniki empiryczne pokazują umiarkowane polepszenie w porównaniu z innymi pesymistycznymi metodami.

e) W nagrodzonej pracy [P5] (NeurIPS 2024 Spotlight), BRO jest złożonym algorytmem łączącym wiele komponentów, co utrudnia zrozumienie, który element odpowiada za daną poprawę wydajności. Ablacje są przeprowadzone, ale złożone interakcje pozostają częściowo niewyjaśnione. Eksperymenty prowadzono głównie na symulacjach komputerowych. Ciekawa byłaby weryfikacja w środowiskach rzeczywistych. Ponadto architektura BroNet wymaga specjalistycznej wiedzy i dokładnego strojenia.

7. Opinia końcowa i wniosek

Doktorant mgr inż. Michał Nauman z powodzeniem rozwiązał istotny i aktualny problem naukowy, jakim jest poprawa efektywności próbkowania algorytmów aktor-krytyk w głębokim uczeniu ze wzmocnieniem. Zaproponowane przez niego oryginalne rozwiązania stanowią znaczące osiągnięcia naukowe, potwierdzone przez recenzentów najważniejszych konferencji w dziedzinie uczenia maszynowego (ICML, NeurIPS, AAAI, IJCAI). Rozprawa dowodzi, że jej Autor posiada głęboką, interdyscyplinarną wiedzę teoretyczną, łączącą teorię procesów decyzyjnych Markowa, teorię optymalizacji, analizę statystyczną i teorię oczekiwanej użyteczności. Ponadto w pełni opanował umiejętność samodzielnego prowadzenia pracy badawczej. Potrafi poprawnie formułować pytania naukowe, przeprowadzać dowody teoretyczne, projektować i implementować algorytmy oraz dokonywać ich eksperymentalnej oceny. Szczególnie znaczące jest to, że doktorant opublikował pracę wyróżnioną nagrodą

Spotlight na NeurIPS 2024 (top 3% zgłoszeń) oraz nawiązał owocną współpracę z wybitnymi badaczami z UC Berkeley (m.in. P. Abbeel, S. Levine).

W świetle powyższych argumentów stwierdzam, że recenzowana rozprawa doktorska mgr Michala Naumana, pt. Sample-Efficient Actor-Critic Algorithms in Reinforcement Learning, spełnia wszystkie wymogi formalne i merytoryczne stawiane rozprawom będącym podstawą nadania stopnia naukowego doktora, określonym w art. 187 ustawy z dnia 20 lipca 2018 r. "Prawo o szkolnictwie wyższym i nauce". Niniejszym pozytywnie opiniuję przedstawioną rozprawę i wnoszę do Przewodniczącego Rady Dyscypliny Naukowej Informatyka Uniwersytetu Warszawskiego o dopuszczenie jej Autora, mgr Michala Naumana, do dalszych etapów postępowania doktorskiego, w tym do publicznej obrony rozprawy. Ponadto, biorąc pod uwagę fakt, że doktorant rozwiązał istotne problemy naukowe, publikując prace w najbardziej prestiżowych konferencjach naukowych z zakresu uczenia maszynowego i sztucznej inteligencji, oraz uzyskał wyróżnienie Spotlight na konferencji NeurIPS 2024, wnioskuję o wyróżnienie niniejszej rozprawy doktorskiej.

Leszek Rutkowski