Rada Naukowa Dyscypliny
INFORMATYKA TECHNICZNA
I TELEKOMUNIKACJA
Sekretariat
Data wpływu 20.11.2024
Numer

dr hab. inż. Jakub Kołota
Instytut Automatyki i Robotyki
Wydział Automatyki Robotyki i Elektrotechniki
Politechnika Poznańska

Poznań, dnia 05.XI.2024

RECENZJA ROZPRAWY DOKTORSKIEJ DLA RADY NAUKOWEJ DYSCYPLINY INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA POLITECHNIKI WARSZAWSKIEJ

Tytuł rozprawy:	**„Leveraging the distribution of collected samples in reinforcement learning"**
Autor rozprawy	**mgr inż. Jakub Łyskawa**
Dyscyplina naukowa:	**informatyka techniczna i telekomunikacja**
Promotor:	**dr hab. inż. Paweł Wawrzyński**

Podstawa formalna sporządzenia recenzji rozprawy doktorskiej: uchwała nr 742/2024 Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja z dnia 17 września 2024 roku w sprawie wyznaczenia recenzentów rozprawy doktorskiej w postępowaniu w sprawie nadania stopnia doktora mgr. inż. Jakubowi Janowi Łyskawie.

I. TEMATYKA, TEZA NAUKOWA I CEL ROZPRAWY

Rozprawa doktorska dotyczy tematyki związanej z zastosowaniem algorytmów uczenia ze wzmocnieniem RL (ang. Reinforcement Learning) off-policy z wykorzystaniem zależności między próbkami wynikających z działania tych algorytmów. Algorytmy uczenia ze wzmocnieniem typu online zakładają, że dane są zbierane ze środowiska w miarę szkolenia agenta, co pozwala na ciągłą ocenę nowych zasad, natomiast podejście off-policy skutkuje gromadzeniem doświadczenia w buforze i ponownym jego wykorzystaniem podczas późniejszego szkolenia. Jest to tematyka dynamicznie rozwijająca się w ciągu ostatnich lat, głównie ze względu na rozwój zasobów sprzętowych oraz wzrost popularności algorytmów sztucznej inteligencji. Autor rozprawy słusznie dostrzega te trendy rozwojowe, zwracając jednak uwagę na ich ograniczenia oraz potencjalne niedoskonałości. Analizie został poddany aktualny stan wiedzy związanej z zastosowaniem algorytmów uczenia ze wzmocnieniem off-policy, natomiast praca prezentuje trzy algorytmy RL, które wykorzystują zależności występujące w zebranych próbkach.

Praca doktorska mgr. inż. Jakuba Łyskawy zawiera treści bezpośrednio związane z dyscypliną naukową informatyka techniczna i telekomunikacja.

Autor sformułował w sposób jawny na str. 11 trzy tezy rozprawy doktorskiej zacytowane poniżej:

Teza 1: „Including the properties of the autocorrelated action noise in the training process improves the sample efficiency in physical control environments."

Teza 2: „Decreasing the expected duration of actions increases the sample efficiency in physical control environments while including the expected action duration in the collected data allows a value-based algorithm to efficiently reuse old experience samples."

Teza 3: „The precision of achieving a subgoal in goal-conditioned HRL may be automatically calculated using a distribution of past distances to subgoals."

Pierwsza z tez odnosi się do poprawy wydajności próby w środowisku fizycznym poprzez włączenie właściwości autokorelowanego szumu akcji do procesu uczenia. W celu jej weryfikacji autor zaproponował nowy algorytm ACERAC (ang. Actor-Critic with Experience Replay and Autocorrelated Actions), ulepszający algorytm ACER (ang. Actor-Critic with Experience Replay) poprzez wprowadzenie autokorelowanego szumu akcji i włączenie właściwości szumu do algorytmu. Weryfikacja empiryczna w symulowanych środowiskach dowiodła słuszności tezy nr 1. Weryfikacja tezy nr 2 odbyła się poprzez opracowanie kolejnego algorytmu SusACER (ang. Actor-Critic with Experience Replay and Sustained Actions) umożliwiającego wykorzystanie doświadczenia zebranego ze zmienną dyskretyzacją czasową. Aby zastosować RL do problemów z natury ciągłych, takich jak sterowanie robotem, należy zdefiniować określoną dyskretyzację czasu. Zaproponowany przez autora rozprawy algorytm łączy zalety różnych ustawień dyskretyzacji czasu. Początkowo działa z rozproszoną dyskretyzacją czasu, aby zmaksymalizować szybkość uczenia się na wczesnych etapach i stopniowo przełącza się na dokładną. Teza nr 3 została zweryfikowana dokonując analizując skutków zmieniającej się dyskretyzacji czasu w środowiskach sterowania robotami.

Rozprawa obejmuje zakresem cykl pięciu publikacji, których wyniki potwierdzają powyższe tezy. Nadrzędny cele badań przedstawionych w rozprawie został osiągnięty – wykazano bowiem, że uwzględnienie własności statystycznych wprowadzonych do doświadczenia zbieranego przez algorytmy RL bez modelu w projektowaniu tych algorytmów może poprawić ich wydajność.

II. ZAWARTOŚĆ MERYTORYCZNA ROZPRAWY

Rozprawa doktorska pt. „Leveraging the distribution of collected samples in reinforcement learning" została napisana w języku angielskim. Zawiera ona bogatą bibliografię liczącą 60 pozycji a 6 rozdziałów stanowiących jej zasadniczą część zostało wzbogaconych o 2 załączniki stanowiące wykaz skrótów oraz pełną treść pięciu współautorskich publikacji.

Pierwszy rozdział ma charakter wprowadzający – przedstawiono w nim podstawowe pojęcia z zakresu tematyki związanej z zastosowaniem algorytmów uczenia ze wzmocnieniem. W tym rozdziale sformułowano także tezy pracy oraz zdefiniowano nadrzędny jej cel. Wylistowano pięć publikacji (P1-P5) stanowiących cykl rozprawy. Rozdział 2 również ma charakter opisowy, gdyż przedstawiono w nim zagadnienia dotyczące procesu decyzyjnego Markova MDP (ang. Markov Decision Process) oraz omówiono podstawowe grupy algorytmów uczenia ze wzmocnieniem (online, offline, model-free, model-based). Kolejne podrozdziały dotyczą konsekwencji wprowadzania kolejnych zmian w uczeniu

RL (wprowadzenie szumu niezależnego od czasu, algorytm ACER). Podrozdział 2.2 opisuje podejście hierarchiczne, które rozwiązuje MDP poprzez rozłożenie go na wiele mniejszych, prostszych problemów. Podrozdział 2.3 w sposób zwarty traktuje o QRL (ang. Quantum Reinforcement Learning). W rozdziale 3 rozprawy autor przedstawia współautorskie podejścia do różnych metod wymuszania podobieństwa kolejnych działań. Metody te są szczegółowo przedstawione w kolejnych podrozdziałach. Następuje w nich odniesienie do cyklu publikacji wskazując największe osiągnięcia.
W P1 i P2 autor wskazuje wyzwanie opracowania algorytmu RL, który uwzględnia autokorelację szumu akcji w procesie uczenia, aby zwiększyć wydajność próbki. W publikacji P2 zdefiniowano na nowo proces autokorelacji. Zastosowano ograniczenie, że rozkład szumu powinien być normalny ze średnią 0 i odchyleniem standardowym σ (hiperparametr). Określono również, że współczynnik korelacji między szumem a jego poprzednią wartością α powinien być stały. W publikacji P1 rozszerzono również środowisko eksperymentalne, aby uwzględnić dokładniejszą dyskretyzację. W tym celu dla każdego z czterech symulowanych środowisk robotycznych uwzględniono również wyższe częstotliwości sterowania. Na podstawie obserwacji wyników opublikowanych w pracy P1 (drobniejsza dyskretyzacja pozwala uzyskać lepsze wyniki, grubsza dyskretyzacja ułatwia naukę) zaprojektowano algorytm SusACER, który zaczyna się od dłuższych akcji i z czasem zmniejsza długość podtrzymania, aby dopasować ją do dyskretyzacji bazowej środowiska. Rozwiązanie zostało przedstawione w publikacji P3. Rozdział 4 opisuje dokonania w obszarze warunkowanych celem algorytmów HRL (ang. Hierarchical Reinforcement Learning), w których parametr SRD (ang. Subgoal Reachability Distance) jest zwykle wybierany ręcznie lub uwzględniany w wyszukiwaniu hiperparametrów, co bezpośrednio zwiększa liczbę hiperparametrów algorytmu. Wybranie zbyt małej wartości SRD może skutkować nierealistycznymi celami, podczas gdy wybranie zbyt dużej wartości może skutkować niedokładną kontrolą. Oba te przypadki utrudniają uczenie całego systemu. W pracy P4 zweryfikowano, iż wartość SRD, która zarówno umożliwia lepszą kontrolę, jak i jest osiągalna przez bieżącą politykę, może być automatycznie wnioskowana z przeszłych wyników. Publikacja P5 ma charakter bardziej przeglądowy. Dokonano w niej przeglądu wpływu technologii kwantowych informacji na badania w różnych obszarach technologii informacyjnych i komunikacyjnych. Jednym z rozważanych obszarów jest RL, gdzie wpływ technologii kwantowych doprowadził do powstania Quantum Reinforcement Learning (QRL). W rozdziale 5 znajduje się podsumowanie, w którym autor zestawia komplementarnie osiągnięcia opisane wcześniej. Ostatnim rozdziałem jest rozdział 6 zawierający dodatkowe osiągnięcia i aktywności autora.

III. OGÓLNA OCENA ROZPRAWY I UWAGI DYSKUSYJNE

Rozprawa doktorska została zredagowana dość starannie, docenić należy także fakt napisania jej w języku angielskim. Pozycje bibliograficzne są dobrane adekwatnie do tematyki pracy, jednakże nie wszystkie zostały opisane w sposób jednolity i kompletny. Doktorant umiejętnie przedstawił własne osiągnięcia zestawiając je ze znajomością tematyki algorytmów uczenia ze wzmocnieniem. W dobie dynamicznego rozwoju algorytmów szerokorozumianej sztucznej inteligencji pojawia się pewien niedosyt w zakresie ukazania skali badań w tej dziedzinie. Autor rozprawy stosunkowo skromnie definiuje treści kolejnych rozdziałów (cały rozdział 4 obejmuje jedną stronę tekstu, rozdział 5 stanowi pół strony tekstu).

Fundament rozprawy obejmuje pięć opublikowanych prac naukowych (2 czasopisma oraz 3 prace konferencyjne) będących w spójnym cyklu monotematycznym:

- **[P1]** Jakub Łyskawa, Paweł Wawrzyński. „ACERAC: Efficient Reinforcement Learning in Fine Time Discretization", IEEE Transactions on Neural Networks and Learning Systems vol. 35(2) (2024).
- **[P2]** Marcin Szulc, Jakub Łyskawa, Paweł Wawrzyński. „A Framework for Reinforcement Learning with Autocorrelated Actions", 27th International Conference on Neural Information Processing (2020).
- **[P3]** Jakub Łyskawa, Paweł Wawrzyński. „Actor-Critic with Variable Time Discretization via Sustained Actions", 30th International Conference on Neural Information Processing (2023).
- **[P4]** Michał Bortkiewicz, Jakub Łyskawa, Paweł Wawrzyński, Mateusz Ostaszewski, Artur Grudkowski, Bartłomiej Sobieski, Tomasz Trzciński „Subgoal Reachability in Goal Conditioned Hierarchical Reinforcement Learning", 16th International Conference on Agents and Artificial Intelligence (2024).
- **[P5]** Bogdan Bednarski, Łukasz Lepak, Jakub Łyskawa, Paweł Pieńczuk, Maciej Rosoł, Ryszard Romaniuk. „Influence of IQT on research in ICT", International Journal of Electronics and Telecommunications, vol. 68, no. 2 (2022).

Praca P1 opublikowana została w najbardziej prestiżowym czasopiśmie spośród wszystkich publikacji autora. Należy zauważyć, że zarówno układ jak i treść pracy osadzonej w rozprawie nie pokrywa się z układem i treścią ostatecznej formy pracy, opublikowanej na stronie wydawcy (https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9834310). Różnice nie są znaczące, jednakże autor powinien zadbać o zaprezentowanie w rozprawie ostatecznej wersji publikacji. Główne osiągnięcie autora (szacowane na 60%) stanowi wprowadzenie stochastycznej zależności między kolejnymi działaniami, umożliwiającej dokładniejszą dyskretyzację czasu w systemach fizycznych. Zaproponowany algorytm umożliwia lepszą eksplorację i co jest z tym związane szybsze uczenie się. Wyniki zaprezentowano dla czterech znanych w robotyce modeli (Ant, HalfCheetah, Hopper oraz Walker2D) z wykorzystaniem platformy symulacyjnej PyBullet. Osiągnięcie to należy ocenić wysoko, tym niemniej nie zostało ono zweryfikowane w warunkach rzeczywistych, o czym wspominają autorzy w podsumowaniu pracy wyznaczając sobie to jako oczywisty krok kolejnych badań. Niestety, późniejsze prace nie zweryfikowały przedstawionych algorytmów z wykorzystaniem rzeczywistych robotów. W pracy P2, w której rozważane są polityki, które generują działania oparte na stanach i losowych elementach autokorelowanych w kolejnych momentach czasowych również odwołano się do korzyści stosowania tych rozwiązań w dziedzinie robotyki. Wnioskowano, iż fizyczna implementacja takich polityk w robotyce, jest mniej problematyczna, ponieważ unika potrząsania robotami. Aby w pełni obronić taką tezę należałoby oprócz symulacji przeprowadzić eksperymenty z wykorzystaniem platformy sprzętowej robota. Podobnie jak w P1, tak i w tej pracy autorzy wskazują kierunek przyszłych prac jako eksperymentalne potwierdzenie skuteczności opracowanych algorytmów. Poniżej zamieszczono ostatnie zdania z sekcji wniosków z obu publikacji P1 oraz P2.

- [P1] „Also, the framework proposed here has been specially designed for applications in robotics. An obvious next step in our research would be to apply it in this area, which is more demanding than simulations."
- [P2] „The framework proposed here is specially designed for applications in robotics. An obvious step of our further research is to apply it in this field, obviously much more demanding than simulations."

Wobec powyższego stwierdzić można, iż autorzy prac doskonale znają wagę i znaczenie potencjalnej eksperymentalnej weryfikacji wyników ich symulacji z wykorzystaniem platformy sprzętowej. W pracy P3 zaproponowano algorytm SusACER, który łączy zalety różnych ustawień dyskretyzacji czasu. Początkowo działa z rozproszoną dyskretyzacją czasu i stopniowo przełącza się na dokładną. Na uznanie zasługuje dominujący udział doktoranta (75%). W pracy zaprezentowano wyniki symulacji używając trzech różnych początkowych wartości oczekiwanych długości akcji E_0. Zastanawiające jest w jaki sposób dokonano wyboru tych konkretnych wartości (2,4 oraz 8)?

W pracy P4 zbadano wpływ długości okna na wydajność ASRD (ang. Adaptive Subgoal Required Distance) dla algorytmu HiTS (ang. Hierarchical reinforcement learning with Timed Subgoals) oraz HAC (ang. Hierarchical Actor-Critic). Podobnie w tej publikacji pojawiają się arbitralne wartości długości okna (50, 500 oraz 1000). W jaki sposób dokonano procesu wyboru badanych wartości? Pytanie to jest szczególnie zasadne wobec wyników zgromadzonych w tabeli nr 2, które ukazują, iż wybór wielkości okna ma kluczowe znaczenie dla wydajności analizowanych algorytmów.

QRL to struktura RL, w której przynajmniej jeden z jej komponentów, polityka i/lub środowisko, jest realizowany przy użyciu systemu kwantowego. Temat ten został podjęty przez doktoranta w pracy P5 prezentowanej w czasopiśmie International Journal of Electronics and Telecommunications (70 pkt). Praca ma sześciu współautorów a udział doktoranta stanowi 1/6 (16.7%). Doktorant zdefiniował jako wkład własny autorstwo sekcji poświęconej kwantowemu uczeniu się przez wzmacnianie (Rozdział V „Quantum reinforcement learning"). Po wnikliwej analizie tegoż rozdziału stwierdzić można, iż doskonała większość treści odnosi się do poniższej publikacji:

Daoyi Dong and Chunlin Chen and Hanxiong Li and Tzyh-Jong Tarn, „Quantum Reinforcement Learning", IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol.38, ISSN=1083-4419, 2008

Doktorant cytuje wskazaną pozycję, jednakże nawet od pracy przeglądowej oczekuje się wartości dodanej, którą nieść powinna publikacja naukowa. Po przeczytaniu całej pracy odnosi się wrażenie chaosu informacyjnego bez myśli przewodniej i wartości twórczej. Publikację oceniam nisko i uważam, iż nie powinna ona widnieć w cyklu publikacyjnym a jedynie stanowić dodatek w dorobku doktoranta. Rozdział II A pracy P5 zaczyna się zdaniem: „There are plenty of CMOS PDKs with accurate room temperature models on the market." Po czym przechodzi się d rozdziału II B, w którym pierwsze zdanie to: „There are plenty of CMOS PDKs with accurate room temperature models on the market." Rozdział VI ma następującą nazwę: „Influence of IQT on biomedicine (MACIEJ'S SECTION)". Po zweryfikowaniu opublikowanej wersji pracy (https://bibliotekanauki.pl) stwierdzono, iż nazwa rozdziału została poprawiona. Świadczy to o dołączonej ponownie nieaktualnej wersji pracy w przedłożonej do recenzji dysertacji.

Wszystkie publikacje doktoranta są współautorskie. Nie powinno to jednak stanowić umocowania do pisania rozprawy w liczbie mnogiej. Autor posługując się liczbą mnogą („przedstawiamy", „pokazujemy", „nasza praca", itd.) odnosi się do dokonania całego zespołu podczas gdy należało skupić się na własnych, podkreślając autorski wkład w każdej wykazanej w cyklu publikacji. To ten zakres prac podlega bezpośredniej ocenie w postępowaniu w sprawie nadania stopnia doktora.
Ostatnia uwaga związana jest z Rozdziałem 6, w którym doktorant wskazuje swoje pozostałe osiągnięcia. Zaprezentowano w nim wystąpienie konferencyjne, jednakże nie ma podanych nawet autorów pracy, brakuje również miejsca konferencji. Doktorant wskazuje projekty, ale nie określa w nich swojej roli. Brak informacji, kiedy przyznano nagrodę dydaktyczną. Nie ma informacji o miejscach wydarzeń z sekcji „social activity", itd. W mojej ocenie to nadmierna niedbałość w prezentowaniu swojego doświadczenia.

IV. UWAGI SZCZEGÓŁOWE

Podczas lektury rozprawy zauważono dość nieliczne usterki mniejszej wagi, które nie wpływają na ogólny odbiór pracy i jej ocenę pod względem merytorycznym. Uwagę zwraca pozostawienie niewypełnionych stron. Należy domniemać, że celem było pozostawienie odstępu pustej strony pomiędzy rozdziałami, jednakże nie jest to konsekwentnie zastosowane powodując niespójność redakcyjną. Na końcach wierszy często znajduje się przedimek „a", który mógłby być przeniesiony do kolejnych wierszy, choć w języku angielskim nie stanowi to błędu. W bibliografii można znaleźć wiele pozycji niedoprecyzowanych. Kilka innych, drobnych wag:

- strona 21 – wzór (2.8) ma zastosowany nieprawidłowy indeks (n),
- strona 21 – niepotrzebne wcięcie tekstu przed wzorem (2.9),
- strona 21 – pozycja cytowana powinna wystąpić przed wzorem (2.10),
- strona 24 – przeniesienie symbolu czasu do nowej linii,
- strona 25 – pojedyncze zdanie stanowiące akapit powinno zostać wchłonięte do treści poprzedniej sekcji,
- strona 25 – podzielona nazwa własna algorytmu pomiędzy liniami tekstu ACERAC,
- strona 34 – brak kropki dla zdefiniowanego osiągnięcia dydaktycznego,
- strona 34 – brak miejscowości prezentowanych wydarzeń,
- Praca P1 oraz P2 ukazuje te same rysunki modeli w platformie symulacyjnej.

Generalnie zaprezentowane w pracy ilustracje, jak i wzory matematyczne zostały sformatowane starannie, są czytelne i prawidłowe. Biorąc pod uwagę fakt napisania pracy w języku angielskim, usterki te są nieliczne, co dobrze świadczy o jakości przygotowania tekstu pod względem edycyjnym. Nie utrudniają one lektury pracy i nie wpływają na jej pozytywną ogólną ocenę.

V. WNIOSKI KOŃCOWE

Opiniowana rozprawa prezentuje ogólną wiedzę teoretyczną doktoranta, jak również umiejętności implementacyjne i programistyczne związane z dyscypliną naukową informatyka techniczna i telekomunikacja, jak też częściowo automatyka, elektronika, elektrotechnika i technologie kosmiczne, co wynika z dobrze pojętej interdyscyplinarności tematyki rozprawy. Autor dysertacji wykazał się umiejętnością rozwiązywania oryginalnego problemu naukowego przedstawionego w recenzowanej

pracy, a uzyskane w niej wyniki były prezentowane we współautorskich publikacjach, m.in. w czasopiśmie IEEE Transactions on Neural Networks and Learning Systems.

Stwierdzam, iż przedstawiona do recenzji rozprawa doktorska mgr. inż. Jakuba Łyskawy pt. „Leveraging the distribution of collected samples in reinforcement learning", której promotorem jest dr hab. inż. Paweł Wawrzyński, spełnia w wystarczającym stopniu wymagania stawiane rozprawom doktorskim przez aktualnie obowiązującą ustawę Prawo o szkolnictwie wyższym i nauce (Dz. U. z 2023 r. poz.742) wraz z odpowiednimi przepisami wprowadzającymi oraz przejściowymi. Wnioskuję o przyjęcie i dopuszczenie do dalszych etapów postępowania doktorskiego, w tym do jej publicznej obrony.

Jakub Kołota