2. Kerr, C. C., Stuart, R. M., Mistry, D., et al. Covasim: An agent-based model of COVID-19 dynamics and interventions. *PLOS Computational Biology*, 2021, vol. 17, no. 7, e1009149. DOI: 10.1371/journal.pcbi.1009149.
3. Vo Hong Thanh, R. & Zunino, R. Tree-based search for Stochastic Simulation Algorithm. In: *Proceedings of the 27th Annual ACM Symposium on Applied Computing (SAC '12)*. Riva del Garda, Italy, 2012, pp. 144–151. DOI: 10.1145/2245276.2232001.
4. Kuryliak, Y. & Emmerich, M. T. M. Towards complexity reduction of large-scale epidemic simulation in two-scale networks. 2025, *CEUR Workshop Proceedings 4004: pp. 327-336.*
5. Du, Z., Yang, Y., Ertem, Z., et al. A review of human mobility: Linking data, models, and real-world applications. *Humanities and Social Sciences Communications*, 2025, vol. 12, no. 1, pp. 1–18. DOI: 10.1057/s41599-025-03160-7

ЗАСТОСУВАННЯ АЛГОРИТМІВ НАВЧАННЯ Q-МЕРЕЖ ДЛЯ ІТЕРАТИВНОЇ ЗАДАЧІ В'ЯЗНЯ

Ткач.Н.В.
Національний університет "Києво-Могилянська академія"
04655, м. Київ, вулиця Сковороди, 2, НаУКМА, Факультет інформатики. тел.(044) 426 60 64.
E-mail: n.tkach@ukma.edu.ua, факс (044) 426 60 64

This study is aimed at showcasing the performance of Deep Q-Networks (DQN) for the Iterated Prisoner's Dilemma (IPD) with a compact episodic state embedding. The agent compresses the interaction context into a fixed-size vector and is trained against deterministic Axelrod strategies. Evaluation of normalized payoff, pairwise cooperation rate of strategies, and the learned behavior of the agent suggests the possibility of efficiently clustering existing strategies by latent learnable features. This may lead to advancements in both game theory and reinforcement learning. The limitations are outlined for future research, including recurrent-based and transformer-based policy-learning networks, stochastic opponents, and comparative analysis to the baseline performance.

Ітераційна дилема в'язня (IPD) є класичною моделлю співпраці[1]. У даній роботі представлено результати агента, навченого на основі алгоритму навчання з підкріпленням глибинних Q-мереж ("deep Q-network")[3], здатного формувати контекстно-залежні рішення за рахунок стислого подання стану. Модель тренувалася проти 72 детерміністичних стратегій у 200-ходових епізодах, реалізованих згідно з турнірами Аксельрода [1] в однойменній бібліотеці.

DQN-агент приймає стислий стан,- 16-вимірний вектор поточної історії гри, що кодує ключові патерни взаємодії. Навчання здійснюється проти детерміністичної популяції у "сліпому" режимі - модель немає інформації про стратегію, проти якої навчається. Функція втрат L(θ) базується на TD-помилці функції оцінювання $V_{\psi}(s)$. Q-функцію[2] ми оцінюємо за допомогою багатошарового перцептрону, який приймає на вхід інформацію про стан s що містить контекст стратегії та останню дію a. Параметрами φ та θ інтерпретуємо як відповідні параметри нейронних мереж які надають оцінку поточного стану, та оцінку стратегії, яку агент має вибрати для мінімізації похибки часової різниці(TD).

$$\min_{\psi}\left(r+\gamma V_{\psi}(s')-V_{\psi}(s)\right)^2 \quad (1)$$

$$y^{\mathrm{DQN}}=r+\gamma \max_{a'} Q_{\theta^-}(s',a') \quad (2)$$

$$L(\theta)=E_{(s,a,r,s')D}\left[\left(y^{DQN}-Q_{\theta}(s,a)\right)^2\right] \quad (3)$$

Внаслідок тренування можна виокремити 4 кластери стратегій (рис.1)які згруповані за очікуваною винагородою та агента та їхнім рівнем кооперації. видно близько чотирьох кластерів стратегій. На рис.2 виведені результати різниці між очікуваним виграшом стратегії та результатами навченого агента.

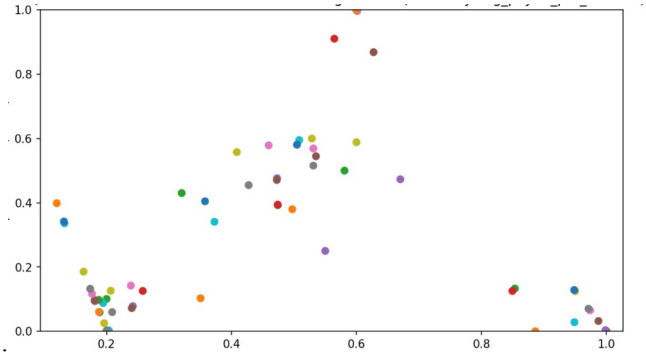

Рис.1Розподіл стратегій у просторі їхньої здатності до кооперації з агентом та зваженого результату агента.

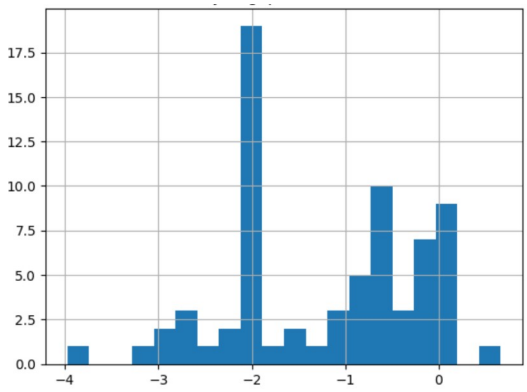

Рис.2 Різниця середнього очікуваного виграшу стратегій відносно навченого агента.

Стисле подання стану дозволяє розрізняти типи опонентів і адаптивно реагувати на їхні патерни поведніки і ефективно відтворювати їх під час фази тестування. Основними обмеженнями цього дослідження є детермінованість популяції, відсутність довготривалої пам’яті та нестаціонарність динаміки навчання. У подальших дослідженнях буде проведено порівняння з рекурентними та трансформер-моделями, введено стохастичних опоненти[4] та популяційний тренінг. У підсумку, можна стверджувати що DQN із компактним поданням показує контекстно-залежну поведінку та виявляє кластерну структуру опонентів, формуючи основу для подальших досліджень кооперації.

ДЖЕРЕЛА

1. Axelrod, Robert, and William D. Hamilton. "The evolution of cooperation." *science* 211.4489 (1981): 1390-1396.
2. Mnih, Volodymyr, et al. "Human-level control through deep reinforcement learning." *nature* 518.7540 (2015): 529-533.
3. Roderick, Melrose, James MacGlashan, and Stefanie Tellex. "Implementing the deep q-network." *arXiv preprint arXiv:1711.07478* (2017).
4. Bertrand, Quentin, et al. "Q-learners Can Provably Collude in the Iterated Prisoner's Dilemma." *arXiv preprint arXiv:2312.08484* (2023).

ЕФЕКТИВНЕ НАВЧАННЯ СТРАТЕГІЙ КЕРУВАННЯ ДЛЯ РОБОТИЗОВАНИХ МАНІПУЛЯЦІЙ ЗА ДОПОМОГОЮ ДИСТИЛЯЦІЇ ЗНАНЬ/EFFICIENT POLICY LEARNING VIA KNOWLEDGE DISTILLATION FOR ROBOTIC MANIPULATION