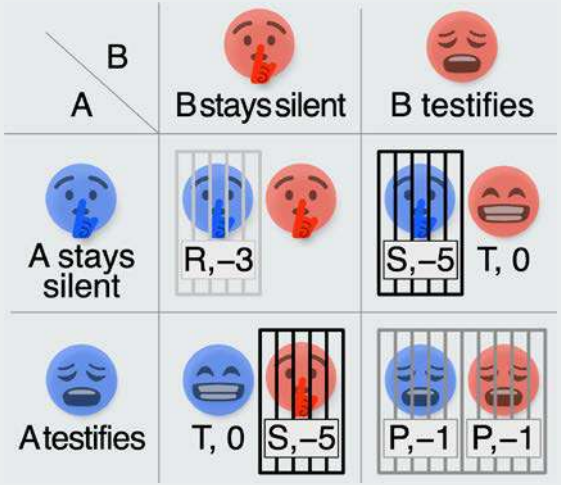
Навчання в ітеративній дилемі в'язня



Виконав: Терентьєв Олександр Андрійович, МП КН-2

Науковий керівник: Ігнатенко Олексій Петрович, доцент, д.н.

Вступ

- Мета дослідження: Розробка та експериментальний аналіз трьох підходів до навчання агентів стратегіям гри ІДВ з метою виявлення їх характеристик, переваг та обмежень у контексті стратегічних ігор

- Завдання дослідження:
    1. Проаналізувати сучасні методи навчання стратегій у стратегічних іграх, зокрема PPO, CMA-ES та Decision Transformer
    2. Розробити уніфіковане програмне середовище для реалізації та тренування агентів за кожним із трьох підходів на основі гри ІДВ
    3. Провести серію експериментів в однакових умовах для об’єктивного порівняння результативності підходів
    4. Проаналізувати отримані результати, порівняти стратегічну поведінку агентів та визначити умови доцільного застосування кожного підходу
    5. Сформулювати висновки та практичні рекомендації щодо застосування розглянутих методів у задачах навчання ігрових стратегій

- Об’єкт дослідження: Процес навчання агентів стратегіям у грі «ітеративна дилема в’язня» засобами різних алгоритмів машинного навчання

Запропонована методологія

Три досліджувані підходи:

1. PPO (Proximal Policy Optimization)
   - Тип: Навчання з підкріпленням
   - Архітектура: 25-вимірний простір спостережень
   - Мережа: Actor-Critic з hidden layers [256, 256, 128]
   - Параметри: learning_rate=3e-4, n_steps=2048

2. CMA-ES Evolution (Memory-One)
   - Тип: Еволюційна оптимізація
   - Стратегія: Memory-One з 5 параметрами (p_cc, p_cd, p_dc, p_dd, initial_action)
   - Алгоритм: CMA-ES, 100 поколінь, популяція 50
   - Sigma: 0.3 (початкове стандартне відхилення)

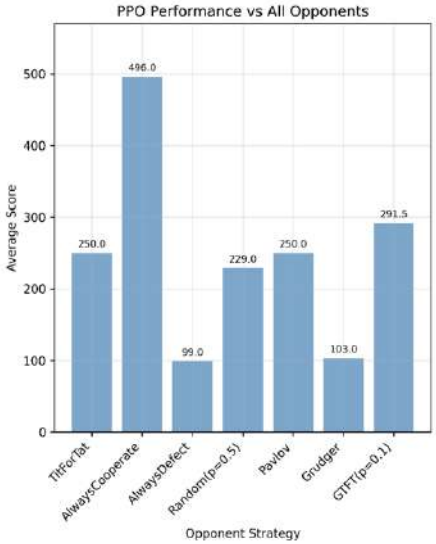
3. Decision Transformer
   - Тип: Контрольоване навчання (supervised learning)
   - Архітектура: 128-dim hidden size, 15-step context length, 3 layers
   - Навчання: Імітаційне навчання на експертних траєкторіях
   - Optimizer: AdamW з learning_rate=3e-4

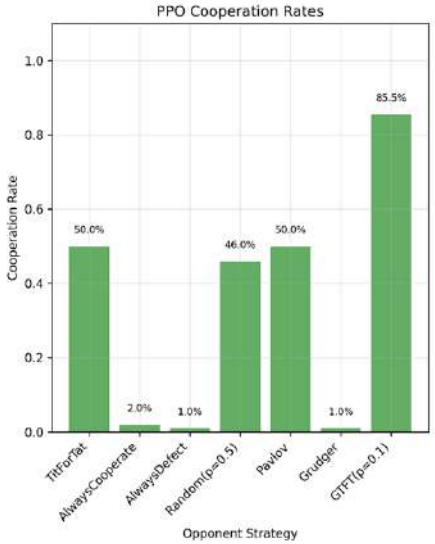
Експериментальне середовище

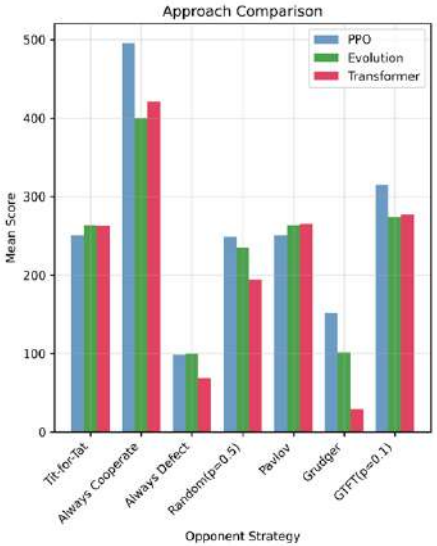
- Уніфіковані параметри експерименту:
  - Довжина гри: 100 раундів
  - Набір опонентів: 7 класичних стратегій
  - Випадкове зерно: 42 (для відтворюваності)
  - Матриця виплат: T=5, R=3, P=1, S=0
- Метрики оцінювання:
  - Середній бал проти кожного опонента
  - Рівень співпраці (частка дій “Cooperate”)
  - Час навчання
  - Стандартне відхилення результатів
- Технічні характеристики:
  - Платформа: Apple M1 Max, 32GB RAM
  - Мова програмування: Python 3.13.2
  - Бібліотеки: Stable-Baselines3, CMA-ES, PyTorch

Результати дослідження – PPO

- Результати PPO:
  - Середній бал: 258.78 (□ 1-е місце)
  - Найкращий результат: 496.0 балів проти Always Cooperate
  - Найгірший результат: 99.0 балів проти Always Defect
  - Середня співпраця: 33.5%
  - Час навчання: 354.6 секунд
- Характеристики підходу:
  - Найвища адаптивність до різних типів опонентів
  - Ефективна експлуатація кооперативних стратегій



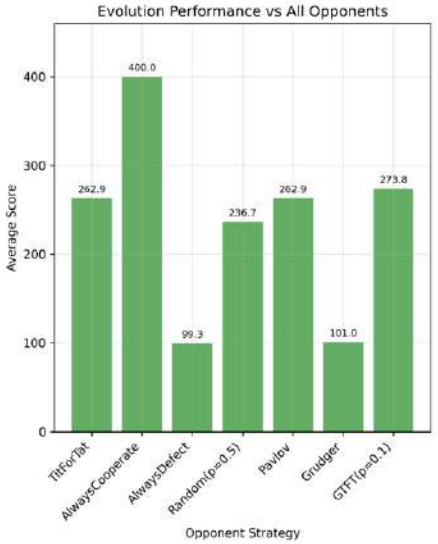


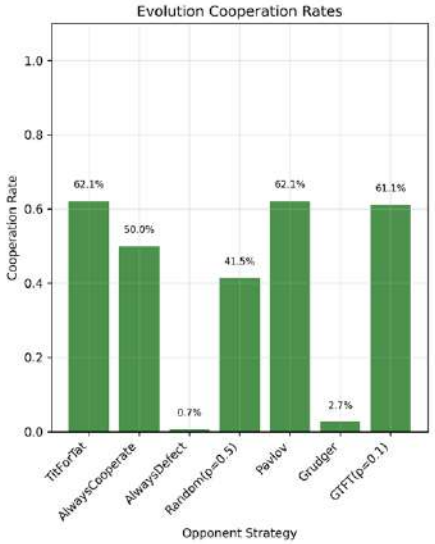


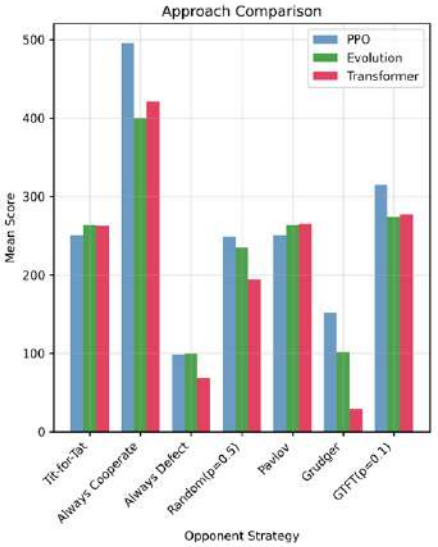
Результати дослідження – еволюційний підхід

- Результати еволюційного підходу:
  - Середній бал: 233.66 (2-е місце)
  - Найкращий результат: 400.0 балів проти Always Cooperate
  - Найгірший результат: 99 балів проти Always Defect
  - Середня співпраця: 40.1%
  - Час навчання: 300.6 секунд (найшвидший)

- Оптимізовані параметри Memory-One стратегії:
  - p_cc = 0.87 (співпраця після взаємної співпраці)
  - p_cd = 0.23 (співпраця після своєї співпраці та зради опонента)
  - p_dc = 0.91 (співпраця після своєї зради та співпраці опонента)
  - p_dd = 0.15 (співпраця після взаємної зради)



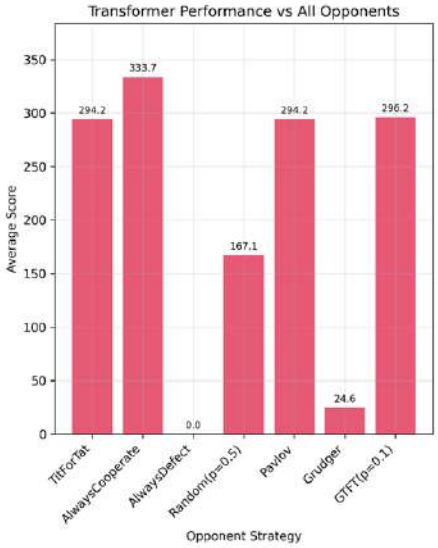


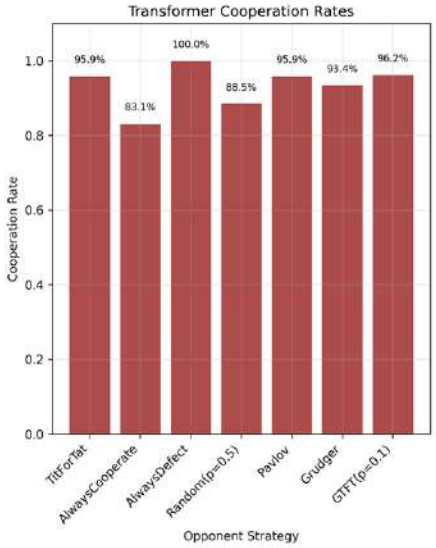


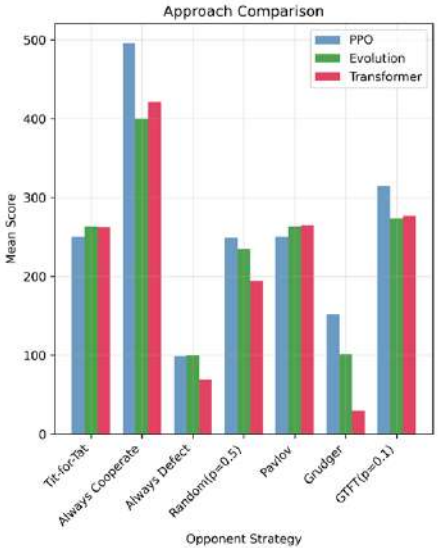
Результати дослідження – Decision Transformer

- Результати Decision Transformer:
  - Середній бал: 217.03 (3-є місце)
  - Найкращий результат: 333.7 балів проти Always Cooperate
  - Найгірший результат: 0 балів проти Always Defect
  - Середня співпраця: 63.0% (найвища)
  - Час навчання: 1115.5 секунд

- Архітектурні параметри:
  - Hidden size: 128 dimensions
  - Context length: 15 кроків
  - Кількість шарів: 3
  - Attention heads: 4



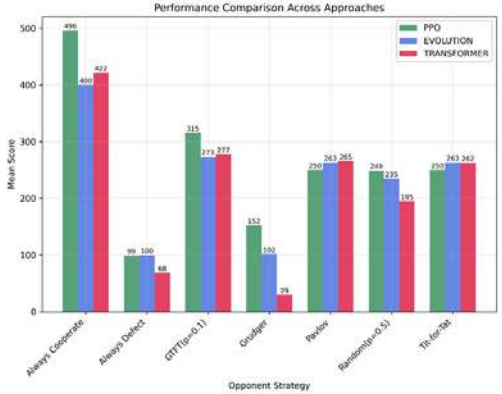


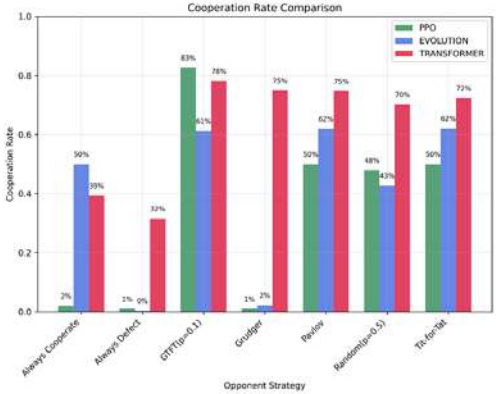


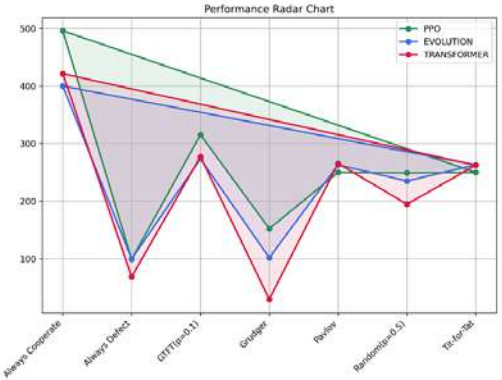
Порівняльний аналіз результатів

- Загальна результативність:
  - PPO: μ=258.78, σ=117.51 – Найвища результативність
  - Evolution: μ=233.66, σ=97.36 – Найкраща стабільність
  - Transformer: μ=217.03, σ=123.99 – Найвища співпраця

- Ключові відмінності в поведінці:
  - PPO: Максимальна експлуатація (496 vs Always Cooperate), мінімальна співпраця проти агресивних опонентів (1-2%)
  - Evolution: Збалансована стратегія зі стабільними результатами (σ=97.36 - найменша варіативність)
  - Transformer: Консистентно кооперативний (60-80% співпраці з більшістю опонентів)







Summary Statistics

Approach	Mean Score	Max Score	Min Score	Std Score
PPO	258.78	496.0	99.0	117.51
EVOLUTION	233.66	400.0	99.62	97.36
TRANSFORMER	217.03	421.52	29.44	123.99

- Методологічні досягнення:
  - Перша систематична порівняльна оцінка трьох парадигм МН для ІДВ в уніфікованих умовах
  - Стандартизований протокол оцінювання з детермінованими параметрами
  - Кількісний аналіз співвідношення “результативність-співпраця-швидкість навчання”
  - Емпіричне встановлення меж застосовності різних типів алгоритмів

- Теоретичні результати:
  - Формалізовано критерії вибору підходу МН для теоретико-ігрових задач
  - Виявлено trade-off між експлуатацією та кооперацією в навчених стратегіях
  - Встановлено кореляції між архітектурними особливостями та стратегічною поведінкою

Практична значимість

Критерії вибору підходу на основі емпіричних результатів:

PPO (Proximal Policy Optimization):
- Застосування: Конкурентні середовища з різноманітними опонентами
- Переваги: Найвища загальна результативність (μ=258.78), адаптивність до контексту
- Обмеження: Низька кооперативна поведінка (33.5% співпраці), висока варіативність стратегії

CMA-ES (Evolution Strategy):
- Застосування: Системи з вимогами до інтерпретованості та швидкості розгортання
- Переваги: Найшвидше навчання (300.6с), повна інтерпретованість стратегії, збалансована співпраця (40.1%)
- Обмеження: Обмежена архітектурна гнучкість (5 параметрів Memory-One)

Decision Transformer:
- Застосування: Кооперативні системи з вимогами до стабільності поведінки
- Переваги: Найвища кооперативність (63.0%), найменша варіативність поведінки
- Обмеження: Вразливість до агресивних стратегій, тривале навчання (1115.5с)

Подальші перспективи

- Безпосередні напрямки розширення дослідження:
  - Тестування на інших 2×2 стратегічних іграх (Chicken Game, Stag Hunt)
  - Валідація результатів із більшою кількістю опонентів-стратегій
  - Порівняння з іншими алгоритмами навчання з підкріпленням (SAC, TD3)

- Технічні покращення методології:
  - Аналіз статистичної значущості відмінностей між підходами
  - Дослідження впливу гіперпараметрів на результативність
  - Розширення метрик оцінювання (стабільність, передбачуваність)

- Практичні застосування результатів:
  - Впровадження в реальні багатоагентні системи
  - Тестування в умовах змінної динаміки середовища
  - Оптимізація для специфічних доменів застосування

Основні висновки

- Результати порівняльного аналізу:
  - Відсутність універсально оптимального підходу – вибір залежить від пріоритетів задачі
  - PPO: найвища результативність (258.78), низька кооперативність (33.5%)
  - CMA-ES: баланс швидкості (300.6с) та інтерпретованості, помірна кооперативність (40.1%)
  - Transformer: найвища кооперативність (63.0%), найповільніше навчання (1115.5с)

- Методологічний внесок:
  - Стандартизований протокол порівняння алгоритмів МН для ІДВ
  - Кількісні критерії вибору підходу

- Обмеження дослідження:
  - Тестування обмежене класичною ІДВ та 7 стратегіями-опонентами
  - Необхідна валідація на інших типах стратегічних ігор

Дякую за увагу!

Репозиторій на GitHub:

https://github.com/alex-teren/learning-in-ipd