Лабораторна робота №1
Тема: Навчання з підкріпленням на основі середовища Frozen Lake бібліотеки Gymnasium.
Методи динамічного програмування та часової різниці (Temporal-Difference)

Підгрупа №3

0. Встановити бібліотеку Gymnasium.
1. Ознайомитись з описом середовища Frozen Lake бібліотеки Gymnasium.
2. Обчислити функцію цінни стану $v_{\pi_1}(s)$ для рівноймовірної (випадкової) стратегії π_1 при параметрі $\gamma = 0{,}75$:
   a. за допомогою ітераційного алгоритму Оцінювання стратегії (Policy Evaluation).
   b. за допомогою розв'язання системи рівнянь Белмана для функції цінни стану відносно невідомих значень $x_i = v_{\pi_1}(s_i)$

   Вивести отримані значення у вигляді матриці або теплової карти. Чи бачите Ви можливі шляхи до покращення рівноймовірної стратегії?
3. Використовуючи знайдені значення функції цінни стану $v_{\pi_1}(s)$ та рівняння Белмана для функції цінни дії-стану, оцінити функцію цінни дії-стану $q_{\pi_1}(s_i, a_j)$.
4. Створити функцію `equiprobable`, результатом якої є номер дії. Дія обирається випадковим чином з множини допустимих дій.
5. Створити функцію `get_episode`, яка приймає у якості аргументу екземпляр середовища, а результатом функції є епізод, тобто список кортежів, кожен з яких зберігає всі характеристики кожного кроку агента (тобто попередній стан, дію, винагороду, поточний стан, значення параметрів terminated та truncated).
   Вибір агентом дії у кожному стані на даному етапі реалізуйте на основі функції `equiprobable`, або, іншими словами, на основі рівноймовірної (випадкової) стратегії π_1.
6. Виконати 100 епізодів за допомогою функції `get_episode`. Виведіть на екран два графіки: винагорода та тривалість епізоду.
7. Реалізувати метод Ітерації цінни (Value Iteration) для знаходження оптимальної стратегії за заданою початковою.
8. Оцінити оптимальну стратегію π_* та функцію цінни стану $v_*(s)$ для заданого середовища за допомогою методу Ітерації цінни, використовуючи в якості початкових параметрів стратегію π_1 та функцію цінни $v_{\pi_1}(s)$. Порівняйте отриману функцію цінни $v_*(s)$ з функцією $v_{\pi_1}(s)$ з завдання №2.
9. Виконати 15 епізодів за допомогою функції `get_episode` зі стратегією π_*. Виведіть на екран два графіки: винагорода та тривалість епізоду. Порівняйте результати з відповідними результатами завдання №6.

10. Використовуючи знайдені значення оптимальної функції ціни стану $v_*(s)$ та рівняння Белмана для функції ціни дії-стану, оцінити оптимальну функцію ціни дії-стану $q_*(s_i, a_j)$. Порівняйте результати з результатами завдання №3.
11. Створити функцію `eps_greedy_policy`, аргументами якої є масив значень функції ціни дії-стану $q(s, a)$ та параметр ε, та результатом є номер дії, обраний з множини номерів допустимих дій допомогою методу ε-жадібної стратегії (ε-greedy policy).
12. Реалізувати метод 𝔼-SARSA (Expected SARSA) для знаходження оптимальної функції ціни дії-стану $q_*(s, a)$ за заданою початковою.
13. Оцінити оптимальну функцію ціни дії-стану $q_*(s, a)$ для заданого середовища за допомогою методу 𝔼-SARSA, використовуючи ε-жадібну стратегію та функцію ціни дії-стану $q_{\pi_1}(s, a)$. Використайте дві стратегії задання значення параметра:
    a. задання постійного значення $\varepsilon \in \{0{,}1, 0{,}5\}$;
    b. зміна значення ε за законом $\varepsilon(k) = \frac{1}{k}$, де $k \in \{1,2,3, \dots, K\}$, K є кількістю епізодів для навчання.

    Порівняйте отримані оцінки функцій $q_*(s, a)$ з функціями $q_{\pi_1}(s, a)$ та $q_*(s, a)$ з завдань №№3, 10. Оберіть кращу оцінку функції $q_*(s, a)$ і відповідне їй значення параметра ε.
14. Створіть на основі функції $q_*(s, a)$ з завдання №13 ε-жадібну стратегію π_2.
15. Виконати 100 епізодів за допомогою функції `get_episode` зі стратегією π_2, отриманою у завданні №14. Виведіть на екран два графіки: винагорода та тривалість епізоду. Порівняйте результати з відповідними результатами завдань №№6, 9.