Міністерство освіти і науки України
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»
Інститут Прикладного Системного Аналізу
Кафедра Математичних Методів Системного Аналізу

Дипломна робота на тему:

Застосування навчання з підкріпленням в задачах оптимального керування

Виконав:
Чаус М.Д., КА-41
Науковий керівник:
к.ф.-м.н.,доцент Яковлева А.П.

Дипломна робота

- **Об’єкт дослідження** - дискретна динамічна система .

- **Предмет дослідження** - застосування алгоритмів навчання з підкріпленням для вирішення задач оптимального керування.

- **Мета роботи** - вирішити задачу оптимального керування для змодельованої динамічної системи за допомогою алгоритмів навчання з підкріпленням.

Актуальність задачі

На сьогоднішній день існує дуже сильна потреба у вирішенні задач оптимального керування. Зазвичай вирішення таких задач полягає у розробці оптимальних контролерів, що була розроблена заздалегідь і вирішена до проведення експеримернту.

Натомість, на практиці дуже часто трапляються здачі, в яких динаміка системи є невідомою, або змінюється в часі. Тому такі системи бажано вирішувати в режимі реального часу без знання про внутрішню динаміку системи.

Постановка задачі

- Треба розробити алгоритми, які вирішать задачу оптимального керування в реальному часі без знання про динаміку системи.

- Провести симуляцію поведінки об’єкта керування і перевірити роботу розроблених алгоритмів.

Лінійно-квадратичний реуглятор

Розглянемо динамічну систему: $x_{k+1} = Ax_k + Bu_k, \quad x_k \in \mathbb{R}^n \quad u_k \in \mathbb{R}^m$

Їй відповідає квадратична функція вартості:

$$V(x_k) = \frac{1}{2}\sum_{i=k}^{\infty} r_i = \frac{1}{2}\sum_{i=k}^{\infty}(x_i^T Q x_i + u_i^T R u_i).$$

Необхідно знайти таке керування, щоб: $u^*(t) = \arg\min_{\substack{u(t) \\ t_0 \le t \le \infty}} V(t_0, x(t_0), u(t))$

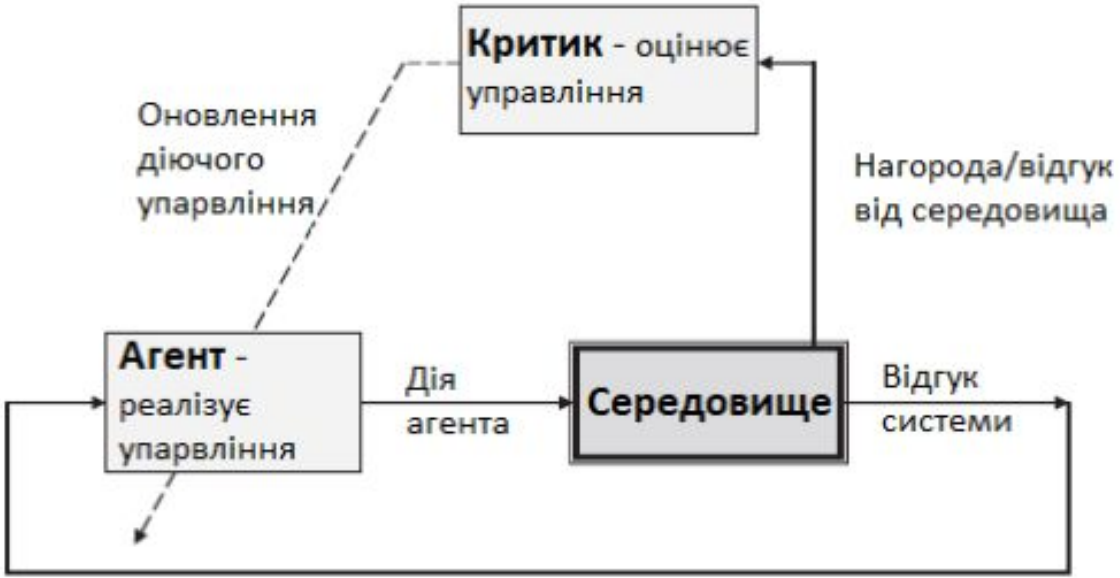
Аналітичний розв’язок має вигляд: $u_k = -Kx_k = -(B^T P B + R)^{-1} B^T P A x_k.$

P - розв’язок алгебраїчного рівнянн Рікатті:

$$A^T P A - P + Q - A^T P B(B^T P B + R)^{-1} B^T P A = 0,$$

Навчання з підкріпленням (Reinforcement learning)

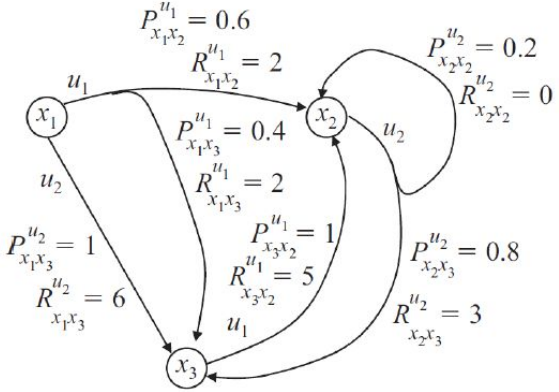
Структура агент-критик.



Марківський процес прийняття рішень

Формально Марківський процес прийняття рішень(МППР) - четвірка (X,U,P,R).

Основним завданням МППР є пошук стратегії $\pi : X \times U \rightarrow [0,1]$, що задає для кожного стану x та переходу u умовну ймовірність $\pi(x,u) = Pr\{u/x\}$, виконання u, якщо МППР знаходиться в стані x.



В нотації МППР для деякої стратегії $\pi(x,u)$ можемо записати функцію вартості

$$V_k^{\pi}(x) = \sum_u \pi(x,u) \sum_{x'} P^u_{xx'} \left[R^u_{xx'} + \gamma V^{\pi}_{k+1}(x') \right].$$

Застосування МППР

Далі застосовуючи принцип оптимальності Беллмана отримуємо наступні вирази для оптимальних значень вартості та керування:

$$V_k^*(x) = \min_u \sum_{x'} P_{xx'}^u \left[R_{xx'}^u + \gamma V_{k+1}^*(x') \right],$$

$$u_k^*(x) = \arg\min_u \sum_{x'} P_{xx'}^u \left[R_{xx'}^u + \gamma V_{k+1}^*(x') \right].$$

які при розгляданні задачі нескінченного горизонту дадуть зручнішу форму, яку вже можна застосовувати на практиці:

$$V^\pi(x) = \sum_u \pi(x,u) \sum_{x'} P_{xx'}^u \left[R_{xx'}^u + \gamma V^\pi(x') \right].$$

$$V^*(x) = \min_u \sum_{x'} P_{xx'}^u \left[R_{xx'}^u + \gamma V^*(x') \right].$$

$$u^*(x) = \arg\min_u \sum_{x'} P_{xx'}^u \left[R_{xx'}^u + \gamma V^*(x') \right].$$

Алгоритм оновлення стратегії

Виберемо початкову стратегію $\pi(x,u)$. Починаючи з j = 0 до сходження робимо послідовно наступні кроки:

а) *Оцінка стратегії (оновлення стратегії):*

$$V_j(x) = \sum_u \pi_j(x,u) \sum_{x'} P^u_{xx'} \left[R^u_{xx'} + \gamma V_j(x')\right]$$

б) *Покращення стратегії (оновлення стратегії):*

$$\pi_{j+1}(x,u) = \arg\min_{\pi(x,u)} \sum_{x'} P^u_{xx'} \left[R^u_{xx'} + \gamma V_j(x')\right]$$

Алгоритм оновлення вартості

Виберемо початкову стратегію $\pi(x,u)$. Починаючи з j = 0 до сходження робимо послідовно наступні кроки:

а) *Оновлення вартості:*

$$V_{j+1}(x) = \sum_{u} \pi_j(x,u) \sum_{x'} P^u_{xx'} \left[R^u_{xx'} + \gamma V_j(x') \right]$$

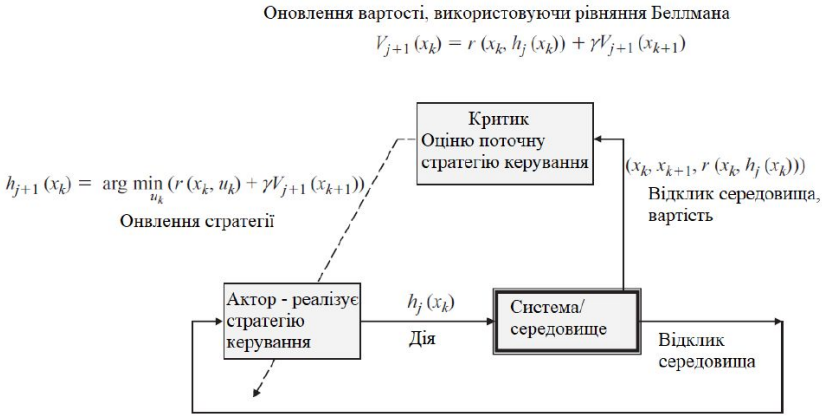
б) *Покращення стратегії:*

$$\pi_{j+1}(x,u) = \arg\min_{\pi(x,u)} \sum_{x'} P^u_{xx'} \left[R^u_{xx'} + \gamma V_j(x') \right], \forall x \in S_j \subseteq X,$$

Застосування нейронних мереж

Для оцінки вартості та стратегії в реальному часі, використовуються дві нейронні мережі, одна з яких апроксимує функцію вартості по останнім спостереженням, а інша апроксимує оптимальне керування.

Ці нейронні мережі відповідають за роботу актора та критика у вище розглянутій моделі.



Моделювання

Модельована система має наступні параметри:

$$A = \begin{bmatrix} -0.0665 & 8 & 0 & 0 \\ 0 & -3.663 & 3.663 & 0 \\ -6.86 & 0 & -13.736 & -13.736 \\ 0.6 & 0 & 0 & 0 \end{bmatrix},$$

$$B = [0, 0, 13.7355, 0]\,.$$

$$Q = I,\ R = I$$

Розв’язки рівняння Ріккаті та оптимального керування, отримані аналітично:

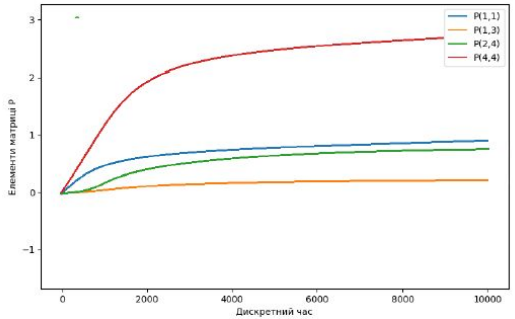
$$\begin{bmatrix} 0.4750 & 0.4766 & 0.0601 & 0.4751 \\ 0.4766 & 0.7831 & 0.1237 & 0.3829 \\ 0.0601 & 0.1237 & 0.0513 & 0.0298 \\ 0.4751 & 0.3829 & 0.0298 & 2.3370 \end{bmatrix}.$$

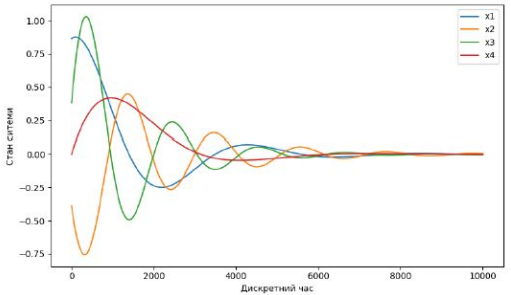
$$K = [0.8267 \quad 1.7003 \quad 0.7049 \quad 0.4142]$$

Результати роботи алгоритму оновлення вартості

Алгоритм оновлення вартості знайшов розв’язок алгебраїчного ріняння Рікатті в реальному часі без знання про динаміку системи з точністю 0.013.

$$P = \begin{bmatrix} 0.4802 & 0.4768 & 0.0603 & 0.4754 \\ 0.4768 & 0.7887 & 0.1239 & 0.3834 \\ 0.0603 & 0.1239 & 0.0567 & 0.0300 \\ 0.4754 & 0.3843 & 0.0300 & 2.3433 \end{bmatrix}.$$

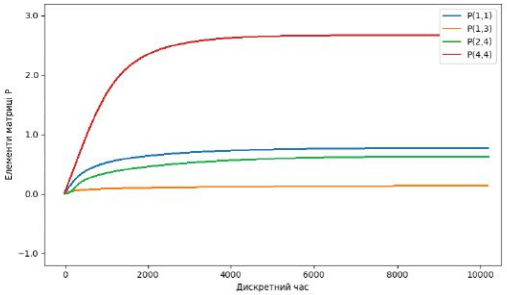


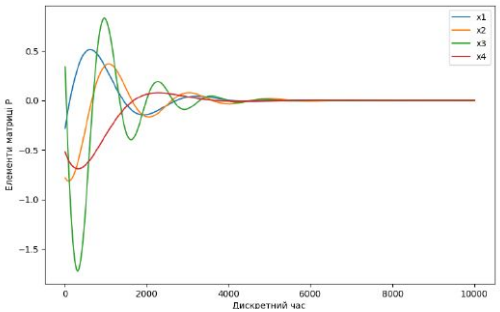


Результати роботи алгоритму оновлення стратегії

Алгоритм оновлення стратегії знайшов розв’язок алгебраїчного ріняння Рікатті в реальному часі без знання про динаміку системи з точністю 0.0005. Причому зійшовся він за меншу кількість ітерацій, ніж алгоритм оновлення вартості.

$$P = \begin{bmatrix} 0.4751 & 0.4766 & 0.0600 & 0.4750 \\ 0.4766 & 0.7830 & 0.1235 & 0.3830 \\ 0.0600 & 0.1235 & 0.0512 & 0.0299 \\ 0.4750 & 0.3830 & 0.299 & 2.3371 \end{bmatrix}$$





Подальші дослідження

- В подальшому планується розробити програмний продукт для розв’язання задач оптимального керування в реальному часі з нелінійною динамікою.

- Застосувати інші методи навчання з підкріпленням та порівняти їх з розробленими.

Дякую за увагу