УДК 004.89+629.7 https://doi.org/10.15407/itm2019.04.029

С. В. ХОРОШИЛОВ, М. О. РЕДЬКА

ІНТЕЛЕКТУАЛЬНЕ КЕРУВАННЯ ОРІЄНТАЦІЄЮ КОСМІЧНИХ АПАРАТІВ ІЗ ВИКОРИСТАННЯМ НАВЧАННЯ З ПІДКРІПЛЕННЯМ

Інститут технічної механіки
Національної академії наук України та Державного космічного агентства України,
вул. Лєшко-Попеля, 15, 49005, Дніпро, Україна; e-mail: skh@ukr.net , mix5236@ukr.net

Метою статті є розробка ефективного алгоритму інтелектуального керування космічними апаратами (КА) на базі методів навчання з підкріпленням (НЗП).

При розробці алгоритму та його дослідженні використано методи теоретичної механіки, теорії автоматичного керування, теорії стійкості, методи машинного навчання та комп'ютерного моделювання. Для підвищення ефективності НЗП використано статистичну модель динаміки, яка базується на понятті гаусових процесів. Така модель, з одного боку, дозволяє використовувати апріорну інформацію про об'єкт керування та має достатню гнучкість, а з іншого – дозволяє охарактеризувати невизначеність у динаміці у вигляді довірчих інтервалів, та може уточнюватися у процесі функціонування КА. У цьому випадку, задача дослідження простору станів-керувань зводиться до отримання таких вимірів, які дозволяють зменшити границі довірчих інтервалів. У якості сигналу підкріплення використано відомий квадратичний критерій, який дозволяє враховувати як вимоги до точності, так і до затрат на керування. Пошук керуючих впливів на базі НЗП виконано із використанням алгоритму ітерацій закону керування. Для реалізації регулятора та оцінювання функції вартості використано апроксиматори у вигляді нейронних мереж. Гарантії стійкості руху КА із врахуванням невизначеності моделі його динаміки отримано з використанням апарату функцій Ляпунова. У якості кандидата функції Ляпунова обрано функцію вартості. Для того щоб спростити перевірку стійкості на базі розглянутої методології, використано припущення про ліпшицеву неперервність динаміки об'єкту керування, що дозволило застосувати метод множників Лагранжа для пошуку керуючих впливів із врахуванням обмежень, сформульованих із використанням верхньої границі невизначеності та ліпшицевих констант динаміки.

Ефективність запропонованого алгоритму ілюструється результатами комп'ютерного моделювання. Запропонований підхід дає можливість розроблювати системи керування, які можуть покращувати свої характеристики по мірі накопичення даних під час функціонування конкретного об'єкту, що дозволяє знизити вимоги до їхніх елементів (сенсорів, виконавчих органів), відмовитись від спеціального стендового обладнання, зменшити терміни та вартість розробки.

Ключові слова: *навчання з підкріпленням, інтелектуальна система керування, космічний апарат, стійкість, модель динаміки.*

Целью статьи является разработка эффективного алгоритма интеллектуального управления космическими аппаратами (КА) на базе методов обучения с подкреплением (ОСП).

При разработке алгоритма и его исследовании использованы методы теоретической механики, теории автоматического управления, теории устойчивости, методы машинного обучения и компьютерного моделирования. Для повышения эффективности ОСП использована статистическая модель динамики, основанная на понятии гауссовых процессов. Такая модель с одной стороны позволяет использовать априорную информацию об объекте управления и обладает достаточной гибкостью, а с другой стороны позволяет охарактеризовать неопределенность в динамике в виде доверительных интервалов и может уточняться в процессе функционирования КА. В этом случае задача исследования пространства состояний-управлений заключается в получении таких измерений, которые позволяют уменьшить границы доверительных интервалов. В качестве сигнала подкрепления использован известный квадратичный критерий, позволяющий учесть, как требования к точности, так к затратам на управление. Поиск управляющих воздействий на базе ОСП выполнен с использованием алгоритма итераций закона управления. Для реализации регулятора и оценивания функции стоимости применены нейросетевые аппроксиматоры. Гарантии устойчивости движения КА с учетом неопределенности модели его динамики получены с использованием аппарата функций Ляпунова. В качестве кандидата функции Ляпунова выбрана функция стоимости. Для того, чтобы упростить проверку устойчивости на базе рассмотренной методологии, использовано допущение о липшицевой непрерывности динамики объекта управления, что позволило применить метод множителей Лагранжа для поиска управляющих воздействий с учетом ограничений, сформулированных с использованием верхней границы неопределенности и липшицевых констант динамики.

Эффективность предложенного алгоритма иллюстрируется результатами компьютерного моделирования. Предложенный подход дает возможность разрабатывать системы управления, которые могут улучшать свои характеристики по мере накапливания данных при функционировании конкретного объекта, что позволяет снизить требования к их элементам (датчикам, исполнительным органам), отказаться от специального стендового оборудования, уменьшить сроки и стоимость разработки.

Ключові слова: *обучение с подкреплением, интеллектуальная система управления, космический аппарат, стойкость, модель динамики.*

The aim of this paper is to develop an effective algorithm for intelligent control of spacecraft based on reinforcement learning (RL) methods.

In the development and analysis of the algorithm, methods of theoretical mechanics, automatic control and stability theories, machine learning, and computer simulation were used. To increase the RL efficiency, a statistical model of spacecraft dynamics based on the concept of Gaussian processes was used. On the one hand, such a model allows one to use a priori information about the plant and is sufficiently flexible, and on the other hand, it characterizes uncertainty in the dynamics in the form of confidence intervals and can be refined during the spacecraft operation. In this case, the problem of control/state space analysis reduces to obtaining such measurements that narrow the confidence intervals. The familiar quadratic criterion, which allows one to take into account both the accuracy requirements and the control cost, was used as the reinforcement signal. An RL-based search for control actions was made using a control law iterative algorithm. To implement the regulator and evaluate the cost function, neural network approximators were used. Spacecraft motion stability guarantees were obtained using the Lyapunov function method with account for the uncertainty in the spacecraft dynamics. The cost function was chosen as a candidate Lyapunov function, To simplify the stability test on the basis of this methodology, the dynamics of the plant was assumed to be Lipschitz continuous, which made it possible to use the Lagrange multiplier method for searching for control actions with account for the constraints formulated using the upper uncertainty bound and Lipschitz dynamics constants.

The efficiency of the proposed algorithm is illustrated by computer simulation results. The approach makes it possible to develop control systems that can improve their performance as data are accumulated during the operation of a specific object, thus allowing one to reduce the requirements for its elements (sensors, actuators), do without special test equipment, and reduce the development time and cost.

Keywords: *reinforcement leaning, intelligent control system, spacecraft, stability, dynamic model.*

Вступ. Система керування орієнтацією та стабілізації (СКОС) грає важливу роль у процесі функціонування сучасних космічних апаратів (КА), тому що від її характеристик багато в чому залежить можливість виконання цільових задач, покладених на КА. При розробці СКОС широко використовуються методи класичної теорії керування [1] та оптимального керування [2], які передбачають наявність точної математичної моделі об’єкта керування (ОК). Однак на практиці усі математичні моделі у тій чи іншій мірі є неточними. Таким чином, параметри об’єкту часто відомі лише приблизно, а його математична модель може бути настільки складною, що це не дозволяє її використати під час синтезу законів керування. Крім цього зовнішні збурення, зазвичай також точно невідомі.

Для керування КА при наявності невизначеності можуть бути використані методи теорії робастного керування [3]. Однак, недоліком такого підходу є те, що робастність алгоритмів керування по відношенню до невизначеності ОК зазвичай досягається за рахунок зниження якості керування. Тому, для забезпечення високих характеристик системи керування, необхідний точний опис невизначеності у тій чи іншій формі, що не завжди можливо.

Іншим напрямком методології керування в умовах невизначеності є теорія адаптивного керування [4], основний принцип якої полягає у отриманні інформації про ОК в процесі його функціонування та використання її для керування. Такий підхід дозволяє підлаштовувати регулятор у процесі функціонування ОК таким чином, щоб забезпечувалась задана точність відстеження деякої оптимальної траєкторії руху, яка розрахована з використанням номінальної моделі. Але, тому що номінальний ОК відрізняється від реального, в цілому таке керування не завжди є оптимальним.

Враховуючи те, що складність та різноманіття задач, які вирішуються за допомогою КА, постійно зростають, використання зазначених вище підходів при розробці СКОС КА призводить до край високих вимог до її елементів (сенсорів, виконавчих органів), необхідності використання спеціального стендового обладнання, високим термінам та вартості розробки.

Розробка СКОС із використанням методів штучного інтелекту має потенціал змінити цю ситуацію. Інтелектуальна система керування може мати можливість покращення характеристик по мірі накопичення даних про особливості функціонування конкретного об'єкту. Такий підхід аналогічний тому, як люди вдосконалюють свої навички по мірі накопичення досвіду.

Серед різноманітних методів штучного інтелекту, в останній час особливий інтерес вчених та практиків направлено на навчання з підкріпленням (НЗП) [7]. Ці методи найбільш близько імітують можливості людини вдосконалювати свою поведінку для досягнення довгострокових цілей по мірі накопичення нового досвіду.

Відомі різноманітні приклади успішного застосування НЗП для вирішення поставлених задач у різних напрямках техніки, наприклад робототехніки [8, 9], транспорту [10, 11], авіації [12].

Однак, інтерес розробників космічної техніки до цього підходу наразі незначний. На нашу думку, це зумовлено низкою причин. По-перше, вважається, що для реалізації такого підходу потрібні значні обчислювальні ресурси, які недоступні на борту КА. Однак це представлення склалось давно, і враховуючи сучасний рівень комп'ютерної техніки та перспективи її розвитку, можна сказати, що доступні на орбіті обчислювальні можливості можуть бути достатні для використання такого підходу. По-друге, відомо, що НЗП властива відносно невисока ефективність навчання. Це призводить до того, що об'єкту необхідно виконати велику кількість спроб, перед тим як він навчиться виконувати необхідну функцію належним чином. Крім цього, у більшості випадків, методологія не забезпечує гарантій досягнення необхідних результатів, до яких звикли розробники космічної техніки.

Таким чином, вдосконалення методів НЗП представляє інтерес із врахуванням специфіки вирішення задач керування КА.

Ціллю статті є розробка ефективного алгоритму інтелектуального керування КА на базі методів навчання із підкріпленням.

Постановка задачі та вхідні дані. Оцінимо можливість використання НЗП для керування кутовим рухом КА на прикладі такої модельної задачі. Припустимо, що на етапі розробки системи керування відома деяка наближена (номінальна) модель динаміки КА, яка відрізняється від реальної як значенням її параметрів, так і деякою динамікою, що не моделюється. З використанням цієї номінальної моделі, синтезуємо базовий алгоритм керування, достатній для виконання КА деякого початкового переліку задач. Далі будемо вважати, що КА, використовуючи цей алгоритм керування, починає функціонувати на орбіті. Потім інтелектуальна система керування виконує послідовно такі дії:

1. Збір даних про особливості динаміки КА;
2. Уточнення моделі динаміки КА з використанням отриманих даних;
3. Покращення алгоритму керування КА з використанням уточненої моделі.

Приведені вище дії повторюються доти, поки забезпечується покращення якості керування. У кінці мають бути отримані такі алгоритми, які максимально наближаються за якістю до оптимального керування, синтезованого з використанням точної математичної моделі ОК.

Математична модель. Для опису кутового руху КА використаємо інерціальну систему координат (ІСК) $O_I x_I y_I z_I$ із початком у центрі мас Землі O_I. Вісь $O_I y_I$ ІСК направлена за віссю обертання Землі, а вісь $O_I z_I$ – у точку весняного рівнодення у задану епоху. Використаємо також зв’язану з КА систему координат (ЗСК) $O_S x_S y_S z_S$ з початком у центрі мас та осями, які співпадають із головними центральними осями інерції апарату.

Рівняння обертального руху абсолютно жорсткого КА можуть бути представлені таким чином:

$$J\dot{\omega} + \omega \times J\omega = \mathrm{M}^d + \mathrm{M}^c, \quad (1)$$

де J – тензор інерції КА; $\omega = [\omega_x, \omega_y, \omega_z]^T$ – вектор абсолютної кутової швидкості КА, заданий проекціями на осі ЗСК; M^d, M^c– вектори сумарного збурюючого та керуючого моментів, відповідно.

У якості параметрів орієнтації використаємо кути Крилова ϑ, φ, ψ (крен, тангаж, рискання). Перехід від ІСК до ЗСК можна зробити послідовністю поворотів (*z*-*y*-*x*) на кути ϑ, φ, ψ. У цьому випадку кінематичні рівняння, які зв’язують вектор абсолютної кутової швидкості КА та похідні кутів орієнтації, можуть бути представлені у такому вигляді:

$$\begin{bmatrix} \dot{\psi} \\ \dot{\varphi} \\ \dot{\vartheta} \end{bmatrix} = \frac{1}{\cos\varphi} \begin{bmatrix} \cos\varphi & \sin\varphi\sin\psi & \sin\varphi\cos\psi \\ 0 & \cos\varphi\cos\psi & -\sin\psi\cos\varphi \\ 0 & \sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} \omega_x \\ \omega_y \\ \omega_z \end{bmatrix}. \quad (2)$$

Далі, при проведенні числових експериментів, будемо вважати, що рівняння (1), (2) точно описують динаміку КА.

Рівняння (1), (2) суттєво нелінійні, але для малих кутових відхилень КА значення похідних кутів орієнтації приблизно рівні $\dot{\psi} \approx \omega_x$, $\dot{\varphi} \approx \omega_y$, $\dot{\vartheta} \approx \omega_z$ та рівняння (1), (2) можуть бути представлені у лінійній формі у вигляді трьох незалежних диференційних рівнянь:

$$J_x\ddot{\psi} = M_x^d + M_x^c, \; J_y\ddot{\varphi} = M_y^d + M_y^c, \; J_z\ddot{\vartheta} = M_z^d + M_z^c, \quad (3)$$

де J_x, J_y, J_z – центральні моменти інерції КА відносно відповідних осей ЗСК; M_x^d, M_y^d, M_z^d і M_x^c, M_y^c, M_z^c – проекції векторів сумарного збурюючого та керуючого моментів на відповідні осі ЗСК.

Вважатимемо, що саме модель (3) відома до виводу КА на орбіту.

Оптимальне керування. Для синтезу регулятора рівняння (3) можна представити у формі простору станів у такому дискретному вигляді:

$$\mathrm{X}_{k+1} = A\mathrm{X}_k + B\mathrm{U}_k, \quad (4)$$

де $\mathrm{X}_k = [\psi, \varphi, \vartheta, \dot{\psi}, \dot{\varphi}, \dot{\vartheta}]^T$, $\mathrm{U}_k = [M_x^c, M_y^c, M_z^c]^T$ – вектори стану та керування на *k*-му такті керування, відповідно.

Матриці стану та керування, які входять у представлення (4), мають вигляд:

$$A=\begin{bmatrix}0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&1&0&0\\0&0&0&0&1&0\\0&0&0&0&0&1\end{bmatrix}\quad B=\begin{bmatrix}J_x^{-1}&0&0\\0&J_y^{-1}&0\\0&0&J_z^{-1}\end{bmatrix}.$$

Для оцінки якості керування використаємо такий квадратичний критерій, який враховує точність керування та затрати на керування:

$$I=\sum_{k=0}^{\infty}\left(X_K^T Q X_K+U_K^T F U_K\right),\qquad(5)$$

де Q та F – вагові матриці.

Для такого критерію і лінійного ОК (4), керування може бути знайдено у формі лінійно-квадратичного регулятора [3]:

$$U_k^L=-KX_k.\qquad(6)$$

Матриця коефіцієнтів підсилення регулятора K знаходиться шляхом розв'язання алгебраїчного рівняння Ріккаті:

$$P=Q+A^T\left(P-PB\left(F+B^T PB\right)^{-1}B^T P\right)A\qquad(7)$$

у такому вигляді:

$$K=\tilde{F}^{-1}B^T P,\ \tilde{F}=F+B^T PB,$$

де P – розв'язок рівняння (7).

Саме керування (6) будемо використовувати далі у якості базового.

Використовуючи метод лінеаризації зворотного зв'язку та вводячи нове керування U_k^*, початкова нелінійна система рівнянь може бути представлена таким чином:

$$X_{k+1}=AX_k+B^*U_k^*,\qquad(8)$$

де $B^*=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}$,

$$U_k^*=-K^*X_k.\qquad(9)$$

Перехід від лінійного керування для системи (8) до нелінійного керування для початкової системи (1), (2) виконується таким чином:

$$U_k^N=\left(JF_kJ^{-1}\right)^{-1}\left(U_k^*-J\dot{F}_k\omega_k^T\right)+\omega_k\times J\omega_k^T,\qquad(10)$$

де

$$F_k=\begin{bmatrix}1&\tan\varphi_k\sin\psi_k&\tan\varphi_k\cos\psi_k\\0&\cos\psi_k&-\sin\psi_k\\0&\sec\varphi_k\sin\psi_k&\sec\varphi_k\cos\psi_k\end{bmatrix},$$

$$\dot{F}_k = \begin{bmatrix} 0 & \left[\sec^2 \varphi_k \sin\psi_k \dot{\varphi}_k + \cos\psi_k \tan\varphi_k \dot{\psi}_k\right] & \left[\sec^2 \cos\psi_k \varphi_k \dot{\varphi}_k - \sin\psi_k \tan\varphi_k \dot{\psi}_k\right] \\ 0 & -\sin\varphi_k \dot{\varphi}_k & -\cos\psi_k \dot{\psi}_k \\ 0 & \left[\sec\varphi_k \left(\sin\psi_k \tan\varphi_k \dot{\varphi}_k + \cos\psi_k \dot{\psi}_k\right)\right] & \left[\sec\varphi_k \left(\cos\psi_k \tan\varphi_k \dot{\varphi}_k - \sin\psi_k \dot{\psi}_k\right)\right] \end{bmatrix}.$$

Синтез керування (9) виконується аналогічно тому, як у випадку з (6).

На рис. 1 – 3 наведено залежності зміни кутів орієнтації від часу при використанні лінійного та нелінійного регуляторів для КА, який має такі центральні моменти інерції: $J_x = 6000$ кг $\cdot$ м^2, $J_y = 4000$ кг $\cdot$ м^2, $J_x = 5000$ кг $\cdot$ м^2. Як видно із цих рисунків, керування лінійним регулятором призводить до значного перерегулювання, і, як наслідок, його якість значно поступається нелінійному. Траєкторії руху КА під керування нелінійного регулятора можна розглядати як еталонні у рамках розглянутої задачі.

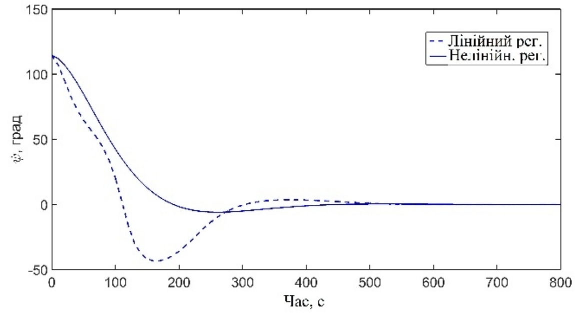

Рис. 1 – Залежність кута крену від часу для лінійного та нелінійного регуляторів

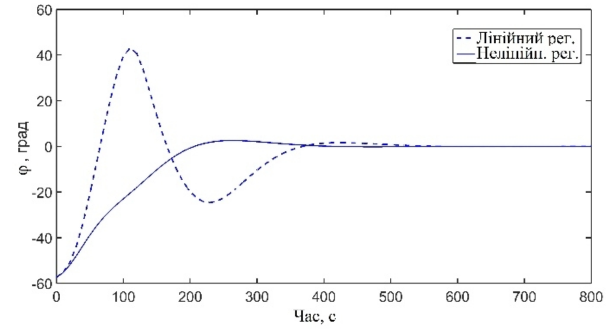

Рис. 2 – Залежність кута тангажу від часу для лінійного та нелінійного регуляторів

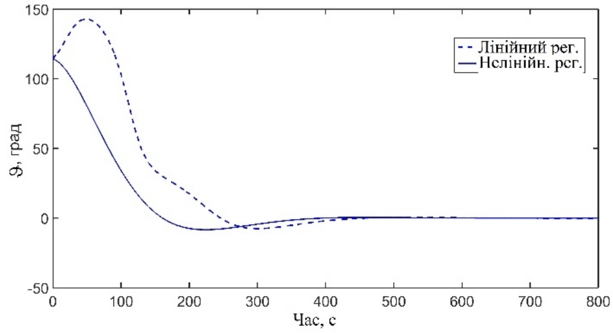

Рис. 3 – Залежність кута рискання від часу для лінійного та нелінійного регуляторів

Навчання з підкріпленням. При вирішенні задач керування із використанням НЗП передбачається, що система керування навчається, аналізуючи результати своїх дій. Ці результати оцінюються за скалярним сигналом (підкріпленням), який отримується від ОК і з яким взаємодіє система керування. Сигнал підкріплення, який можна трактувати як вартість, дозволяє інтелектуальній системі керування змінювати свої алгоритми керування, враховуючі досягнення довгострокової цілі.

Загальний алгоритм НЗП, наведений на рис. 4, включає такі дії:

1) у момент часу t_k ОК знаходиться у стані X_k;

2) у цьому стані система керування обирає один із можливих керуючих впливів (дій) U_k;

3) система керування виконує цю дію, що призводить до переходу ОК у новий стан X_{k+1} і отримання підкріплення R_k;

4) виконується перехід до пункту 2 із врахуванням отриманого підкріплення, або, якщо новий стан є кінцевим, то виконується завершення алгоритму.

Нехай χ – множина станів, а A – множина керуючих впливів. Підкріплення R_k є наслідком дії U_k, обраної у стані X_k. Сигнал підкріплення являє собою функцію, яка залежить від вектору, визначеного у просторі $\chi \times A$.

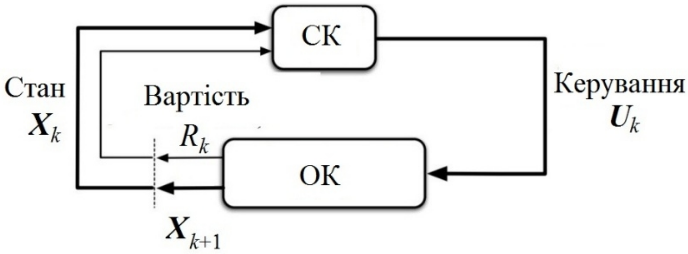

Рис. 4 – Схема навчання з підкріпленням

Система керування обирає дії таким чином, щоб мінімізувати сумарну вартість, яка визначається наступним чином:

$$G_k = R_k + \gamma R_{k+1} + \gamma^2 R_{k+2} + \ldots = \sum_{i=0}^{\infty} \gamma^i R_{k+i}, \quad 0 \le \gamma \le 1 .$$

Коефіцієнт знецінення γ задає ступінь важливості прогнозних значень вартості у майбутньому при виборі керуючих впливів.

Одним із ключових понять ОЗП є функція вартості. Нехай у кожному стані X_k система керування формує керуючу дію згідно до визначеної стратегії π:

$$U_k = \pi(X_k),$$

тоді функція вартості дозволяє визначити сумарну вартість дій при русі із початкового стану X_k і виборі керуючих дій згідно до стратегії π. Цю функцію можна представити таким чином:

$$V^{\pi}(X_k) = R_k(X_k, U_k) + \gamma R_{k+1}(X_{k+1}, U_{k+1}) + \ldots =$$

$$= \sum_{i=0}^{\infty} \gamma^k R_{k+i}(X_{k+i}, U_{k+i}) = R_k(X_k, U_k) + \gamma V^{\pi}(X_{k+1}).$$

Модель для НЗП. Алгоритм НЗП може бути реалізовано із використанням моделі ОК. Така модель повинна описувати перехід ОК із початкового стану X_k під дією керування U_k у наступний стан X_{k+1} таким чином:

$$X_{k+1} = f(X_k, U_k). \tag{11}$$

Цю модель доречно представити у такому вигляді:

$$X_{k+1} = h(X_k, U_k) + g(X_k, U_k), \tag{12}$$

де $h(X_k, U_k)$ – номінальна модель; $g(X_k, U_k)$ – невизначеність.

Особливістю НЗП із використанням моделі є те, що при навчанні використовується інформація не про реальні переходи ОК під дією керуючих впливів, а аналогічні дані, отримані за допомогою моделі. При такому підході якість отриманих результатів визначається точністю використаної при навчанні моделі ОК. У зв’язку з цим, для вирішення розглянутої задачі, використаємо таку модель ОК, яка має потенціал для уточнення. Необхідно запропонувати алгоритми, які дозволяють зменшувати невизначеність $g(X_k, U_k)$ моделі по мірі накопичення даних про особливості функціонування ОК.

Такі моделі можуть бути отримані із використанням різних підходів, наприклад методів механіки. У такому випадку, структура моделі буде фіксованою, а її параметри можуть бути уточнені з використанням методів параметричної ідентифікації [13]. Однак для нашої задачі такий підхід не є доречним, тому що у нашому випадку модель має як параметричну, так і структурну невизначеність.

Великий потенціал для опису різних процесів за експериментальними даними мають моделі, які побудовані на базі нейронних мереж. Нажаль, такий підхід потребує дуже великої кількості даних для забезпечення якісних результатів, що ускладнює їх використання у випадку КА.

Враховуючи наведені вище складності, у цій роботі використана статистична модель, яка базується на понятті гаусових процесів [14]. Такий підхід дозволяє отримати апостеріорне розподілення функції $f(X_k, U_k)$ за наявними даними із використанням непараметричної байєсової регресії:

$$p(f) = GP(\mu, k).$$

Гаусовий процес *GP* повністю визначається функцією математичного очікування μ та додатньою коваріаційною функцією k, яка також називається ядром.

У якості функції математичного очікування доцільно обрати номінальну модель ОК:

$$\mu_k = h(X_k, U_k).$$

У якості ядра зазвичай обирають стандартні коваріаційні функції (рис. 5), які найбільш повно задовольняють особливостям процесу, що розглядається.

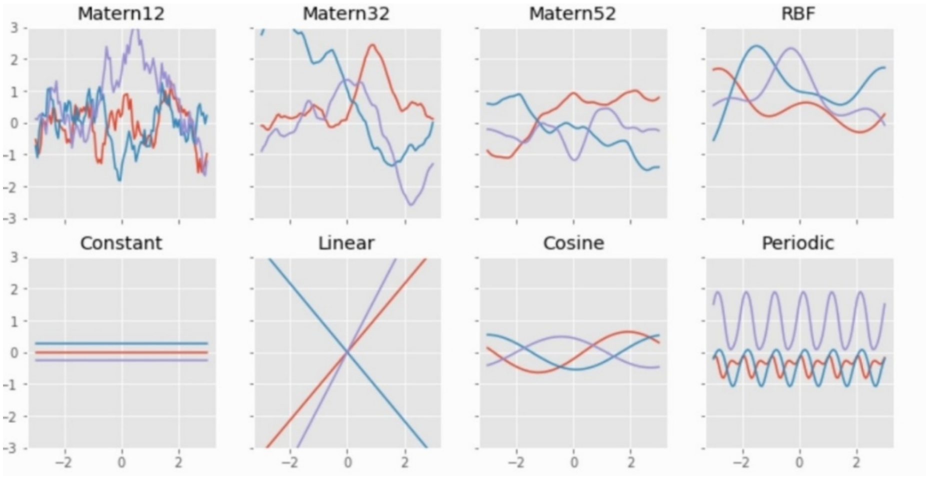

Рис. 5 – Стандартні коваріаційні функції

Під даними будемо розуміти отримані у процесі функціонування КА зашумлені виміри вигляду:

$$Y_k = \hat{f}(Z_k) = f(Z_k) + \varepsilon,\ Z_k = (X_k, U_k),\ \varepsilon \to N\left(0, \sigma_\varepsilon^2\right).$$

Нехай у нас є n навчальних наборів вхідних даних $\overline{Z} = \left[Z_1, Z_2, \ldots, Z_n\right]$ та відповідні їм виміри виходу $\overline{Y} = \left[Y_1, Y_2, \ldots, Y_n\right]$.

Апостеріорне розподілення $p(f_* | Z_*)$ функції $f(Z_k)$ для довільного, але відомого тестового входу Z_* також є гаусовим. Математичне очікування та дисперсія цього розподілення визначаються таким чином:

$$\mu(Z_*) = k_*^T\left(K + \sigma_\varepsilon^2 I\right)^{-1}\overline{Y},\ \sigma^2(Z_*) = k_{**} - k_*^T\left(K + \sigma_\varepsilon^2 I\right)^{-1}k_* + \sigma_\varepsilon^2,$$

де $k_* = K\left(\overline{Z}, Z_*\right),\ k_{**} = K(Z_*, Z_*),\ K_{ij} = k\left(Z_i, Z_j\right)$.

На рис. 6 наведено опис динаміки ОК із використанням GP. Тут світло-сірим кольором показано область значень, куди з високою ймовірністю потрапляє реальний процес. На лівому рисунку показана ця область (довірчі інтервали) перед початком отримання вимірів про динаміку ОК. Два наступних рисунки показують, як виміри, що надходять (позначені хрестиком), дозволяють послідовно звузити довірчі інтервали, тим самим зменшивши невизначеність у динаміці ОК.

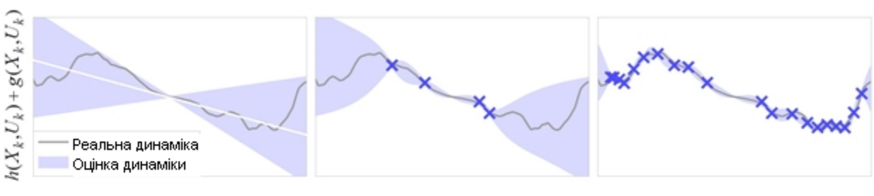

Рис. 6 – Опис ОК за допомогою гаусових процесів

Архітектура «Виконавець – Критик». Існують різні алгоритми знаходження оптимального керування із використанням НЗП. У цей роботі обрано алгоритм ітерацій закону керування, який має кращу збіжність у порівнянні з іншими алгоритмами, однак він програє їм з точки зору ефективності навчання (потребує більше даних для навчання). Враховуючи те, що при навчанні використовується модель, а не реальні переходи ОК, цей фактор не є визначальним.

Суть цього алгоритму полягає у позмінному уточненні функції вартості і закону керування і включає такі кроки:

1. Обирається початковий закон керування π;
2. Оцінюється функція вартості V^{π} для цього закону керування;
3. Виконується деяке число ітерацій з уточнення закону керування, виходячи із мінімізації такої цільової функції: $\pi(\mathrm{X}) \leftarrow \arg\min_{\mathrm{U}} \left[R(\mathrm{X}_k, \mathrm{U}_k) + \gamma V^{\pi}(\mathrm{X}_k) \right].$

Кроки 2 та 3 повторюються доти, поки не буде отриманий оптимальний закон керування $®^{*}$ і відповідна йому функція вартості $V^{®^{*}}$.

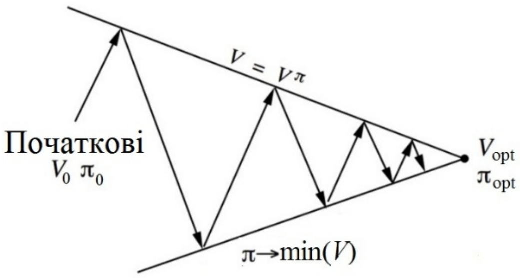

Рис. 7 – Алгоритм ітерацій закону керування

Такий алгоритм може бути реалізовано із використанням двох модулів – критика та виконавця. У цьому випадку, критик формує на виході оцінки функції вартості, а виконавець – керуючі впливи.

Критика та виконавця реалізовано у формі багатошарових нейронних мереж із прямим поширенням сигналів, які апроксимують відповідно функцію вартості та закон керування:

$$V_{\eta}^{\pi}(\mathrm{X}),\ \pi_{\theta}(\mathrm{X}),$$

де η, θ – вектори параметрів критика та виконавця відповідно.

Для навчання критика використано метод скінчених різниць (СР) [7], оснований на мінімізації помилки СР, яка обчислюється таким чином:

$$\delta_k = R_k + \gamma V^{\pi}(\mathrm{X}_{k+1}) - V^{\pi}(\mathrm{X}_k).$$

Враховуючи це, цільова функція критика приймає такий вигляд:

$$V_{\eta}^{\pi}(\mathrm{X}) \leftarrow \arg\min_{V_{\eta}^{\pi}} \left[\mathrm{R}_k + \gamma V^{\pi}(\mathrm{X}_{k+1}) - V^{\pi}(\mathrm{X}_k) \right].$$

Цільова функція виконавця із використанням оцінок критика сформована таким чином:

$$\mathrm{U} = \pi_{\theta}(\mathrm{X}) \leftarrow \arg\min_{\pi_{\theta}} \left[R(\mathrm{X}_k, \pi_{\theta}) + \gamma V_{\eta}^{\pi}(\mathrm{X}_{k+1}) \right]. \qquad (13)$$

Аналіз стійкості та робастність системи. Оптимізація закону керування із використання НЗП часто призводить до того, що замкнутий контур системи керування знаходиться на межі області стійкості. Враховуючи те, що використана у цій роботі для навчання модель ОК містить невизначеність, керований рух реального КА може стати нестійким. Тому критерій (13) повинен бути доповнений умовами, які враховують невизначеність моделі.

У розглянутому у цій статті прикладі, КА починає функціонувати з регулятором, синтезованим із використанням лінійної моделі, яка адекватно описує динаміку тільки для малих відхилень компонент вектору стану від нульового положення, і, таким чином, стійкість замкнутого контуру системи керування забезпечується лише для деякої підмножини простору станів. Будь-які траєкторії керованого руху ОК, які беруть початок всередині області атракції, у кінці кінців, збігаються до цільового стану.

У роботі [15] для отримання гарантій стійкості авторами робляться деякі припущення про динаміку системи. Так, передбачається, що динаміка системи ліпшицево неперервна. Слідуючи результатам цієї роботи, також будемо рахувати, що функції $h(X_k, U_k)$ та $g(X_k, U_k)$, які описують динаміку ОК, та керування $\pi(X_k)$ відповідно L_h, L_g та L_π – ліпшицево неперервні. Тут L_h, L_g та L_π – відповідні ліпшицеві константи.

Крім цього, будемо рахувати, що наша статистична модель динаміки задовольняє такій вимозі:

$$\left\| f(Z_*) - \mu(Z_*) \right\|_1 \leq \beta\sigma(Z_*).$$

Тут масштабуючий коефіцієнт β обирається таким чином, щоб забезпечити із високою ймовірністю потрапляння значень функції $f(Z_*)$ у обраний довірчій інтервал.

Метод функцій Ляпунова широко використовується у теорії керування для дослідження стійкості. Для випадку, коли динаміка системи ліпшицево неперервна, функція Ляпунова v є квазіопуклою в межах області атракції. Ця особливість дозволяє замінити (для перевірки стійкості) вимогу від’ємності похідної від функції Ляпунова вимогою її убування на одному кроці:

$$v(f(X_k, \pi_\theta(X_k))) < v(X_k). \tag{14}$$

Використана у цій роботі функція вартості (5) строго додатна на усій розглянутій області змінення вектору стану, крім нуля. Враховуючи це, у якості кандидата функції Ляпунова оберемо функцію вартості:

$$v(X_k) = V^\pi.$$

Крім цього, слід зазначити, що використана нами модель динаміки містить невизначеність. Тому, оцінки функції Ляпунова також будуть мати деяку невизначеність. Для того щоб врахувати це, умову (14) можна представити у такому вигляді:

$$u_*(X_k, \pi_\theta(X_k)) < v(X_k) - L_{\Delta v}\tau, \tag{15}$$

де u_* – верхня границя змінення функції V; τ – шаг дискретизації простору станів; $L_{\Delta V}$ – ліпшицева константа, яка визначається таким чином:

$$L_{\Delta V} = L_V L_f \left(L_\pi + 1\right) + L_V .$$

Цільова функція виконавця (13) може бути доповнена вимогою робастної стійкості (15) за допомогою методу множників Лагранжа таким чином:

$$\pi \leftarrow \arg\min_{\pi_\theta} \begin{bmatrix} R(\mathrm{X}_k, \pi_\theta(\mathrm{X}_k)) + \\ \gamma V_\eta^\pi(\mu_*(\mathrm{X}_k, \pi_\theta(\mathrm{X}_k))) + \lambda(u_*(\mathrm{X}_k, \pi_\theta(\mathrm{X}_k)) - V(\mathrm{X}_k) + L_{\Delta V}\tau) \end{bmatrix}, \quad (16)$$

де λ – множник Лагранжа.

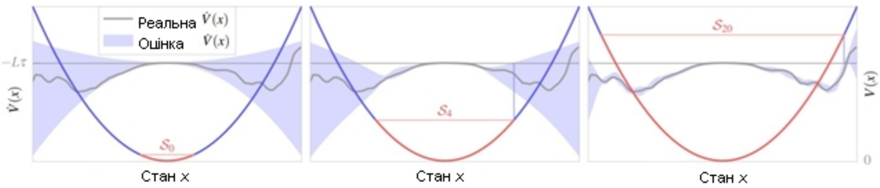

Рис. 8 – Перевірка приналежності вектору стану області атракції

Алгоритм ефективного НЗП. На базі наведеної вище методології, алгоритм ефективного НЗП для СКОС КА може бути представлено таким чином:

1. Вибір детермінованої моделі динаміки КА $h(\mathrm{X}_k, \mathrm{U}_k)$.
2. Синтез базового регулятора $\pi(\mathrm{X}_k)$ із використанням моделі $h(\mathrm{X}_k, \mathrm{U}_k)$.
3. Навчання виконавця $\pi_\theta(\mathrm{X})$.
4. Навчання критика $V_\eta^\pi(\mathrm{X})$.

Початок циклу

5. Збір даних про особливості динаміки КА із використанням керування $\pi_\theta(\mathrm{X})$.
6. Уточнення статистичної моделі $g(\mathrm{X}_k, \mathrm{U}_k)$.

Початок циклу

7. Покращення керування $\pi_\theta(\mathrm{X})$.
8. Уточнення критика $V_\eta^\pi(\mathrm{X})$.

Кінець циклу

Кінець циклу

Числові експерименти. При проведенні числових експериментів передбачалося, що спочатку відома лише лінійна модель КА (3). Точність визначення її параметрів $\pm 20\%$ від її істинних значень, наведених у розділі «Оптимальне керування». Обрано наступні структури нейронних мереж:

критика: число прихованих шарів – 4, число нейронів у прихованому шарі – 64, функції активації – лінійний випрямний елемент (ReLU);

виконавця: число прихованих шарів – 3, число нейронів у прихованому шарі – 64, функції активації ReLU (всюди крім виходу) і Tanh (на виході).

На рис. 9 показано зміну цільової функції виконавця при його навчанні лінійному керуванню (6) із використанням методу навчання із вчителем.

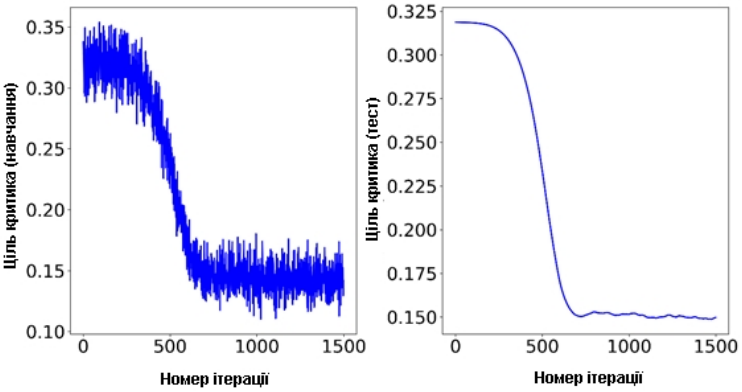

Рис. 9 – Навчання виконавця базовому керуванню

На рис. 10 наведено графіки цільових функцій критика та виконавця у процесі реалізації алгоритму ітерацій закону керування. При навчанні використано коефіцієнт знецінення $\gamma = 0{,}994$.

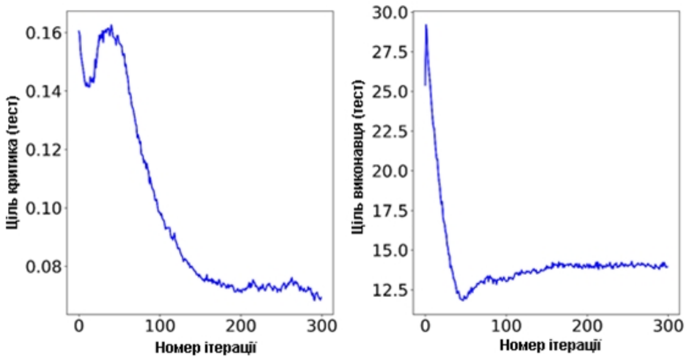

Рис. 10 – Ітерації закону керування

Як видно з рис. 11 – рис. 13, запропонований алгоритм ОЗП дозволяє суттєво покращити базовий лінійний регулятор. Для цього розрахункового випадку, сумарні вартості для лінійного та інтелектуального регуляторів склали 882,491 і 720,718, відповідно. Однак результати інтелектуального регулятора дещо поступаються результатам нелінійного регулятора із розділу «Оптимальне керування», який отримано з використанням точної моделі ОК. Це пояснюється такими факторами. По-перше, у нашому випадку, при НЗП передбачається деяка залишкова невизначеність моделі динаміки КА, що призводить до необхідності отримання робастного регулятора, який, як правило, консервативний. По-друге, при навчанні критика з використанням метода скінчених різниць, було обрано коефіцієнт знецінення менше одиниці. Це, як правило, необхідно для забезпечення збіжності використаного алгоритму при використанні апроксиматорів (нейронних мереж).

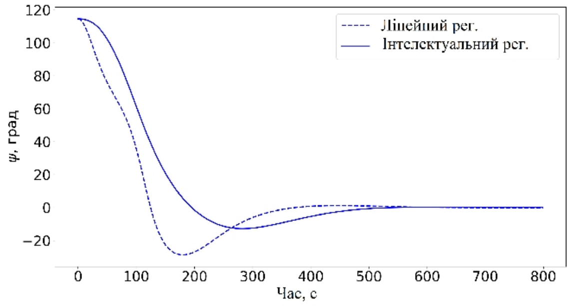

Рис. 11 – Залежність кута крену від часу для лінійного та інтелектуального регуляторів

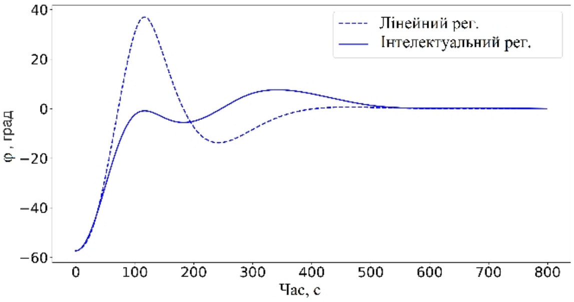

Рис. 12 – Залежність кута тангажу від часу для лінійного та інтелектуального регуляторів

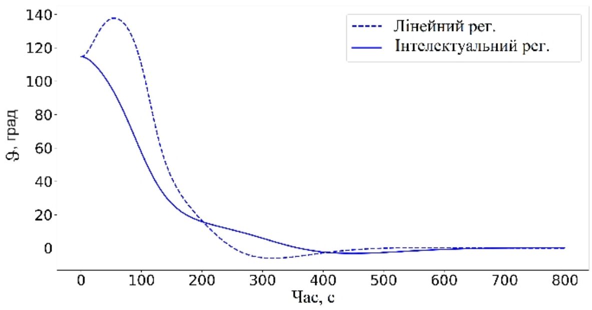

Рис. 13 – Залежність кута рискання від часу для лінійного та інтелектуального регуляторів

Висновки. У статті продемонстрована можливість покращення якості керування КА у процесі його функціонування із використанням навчання із підкріпленням. Для підвищення ефективності навчання, використано модель, яка базується на понятті гаусових процесів. Застосування апарата функцій Ляпунова дозволяє гарантувати стійкість керованого руху при використанні таких моделей.

Запропонований метод дає можливість розробляти системи керування, які можуть покращувати свої характеристики по мірі накопичення даних при функціонуванні конкретного об’єкту. Методологія дозволяє знизити вимоги до елементів систем керування (сенсорів, виконавчих органів), відмовитись від спеціального стендового обладнання, знизити терміни та вартість розробки.

Представляється можливим покращити запропонований алгоритм шляхом використання іншого методу навчання критика, наприклад методу навчання Монте-Карло [7], що може бути напрямком подальших досліджень.

Дослідження проведені за рахунок фінансування за бюджетною програмою "Підтримка розвитку пріоритетних напрямків наукових досліджень" (КПКВК 6541230).

1. *Бесекерский В. А., Попов Е. П.* Теория систем автоматического управления. 4-е изд СПб.: Профессия, 2003. 768 с.
2. *Лейтман Дж.* Введение в теорию оптимального управления. Москва: Наука, 1968. 192 с.
3. *Zhou K., Doyle J.C., Glover K.* Robust and optimal Control. NJ : Prentice-Hall, 1996. 596 p.
4. *Alpatov A., Khoroshylov S., Bombardelli C.* Relative Control of an Ion Beam Shepherd Satellite Using the Impulse Compensation Thruster. Acta Astronautica. 2018. Vol. 151. P. 543–554. https://doi.org/10.1016/j.actaastro.2018.06.056
5. *Astrom K. J., Wittenmark B.* Adaptive Control. MA : Addison-Wesley, 1995. 580 p.
6. *Хорошилов С. В.* Управление ориентацией солнечной электростанции космического базирования с использованием наблюдателя для расширенного вектора состояния. Техническая механика. 2011. Вып. 3. С.117–125.
7. *Sutton R. S., Barto A. G.* Reinforcement learning: an introduction. MIT press, 1998. 338 p.
8. *Gullapalli V.* Skillful control under uncertainty via direct reinforcement learning. Reinforcement Learning and Robotics. 1995. Vol. 15(4). P. 237–246. https://doi.org/10.1016/0921-8890(95)00006-2
9. *Kober J., Bagnell J. A., and Peters J.* Reinforcement learning in robotics: A survey. International Journal of Robotic Research. 2013. Vol. 32(11). P. 1238–1274. https://doi.org/10.1177/0278364913495721
10. *Theodorou E., Buchli J., Schaal S.* Reinforcement learning of motor skills in high dimensions. In International Conference on Robotics and Automation (ICRA), 2010. P. 2397–2403. https://doi.org/10.1109/ROBOT.2010.5509336
11. *Endo G., Morimoto J., Matsubara T., Nakanishi J., Cheng G.* Learning CPG-based biped locomotion with a policy gradient method: Application to a humanoid robot. International Journal of Robotic Research. 2008. Vol. 27(2). P. 213–228. https://doi.org/10.1177/0278364907084980
12. *Ng A. Y., Kim H. J., Jordan M. I., Sastry S.* Inverted autonomous helicopter flight via reinforcement learning. In International Symposium on Experimental Robotics, 2004. P. 363–372. https://doi.org/10.1007/11552246_35
13. *Juang J.-N.* Applied System Identification. N.J: Prentice Hall, Upper Saddle River, 1994. 394 p.
14. *Seeger M.* Gaussian Processes for Machine Learning. International Journal of Neural Systems. 2004. Vol. 14 (2).P. 69–104. https://doi.org/10.1142/S0129065704001899
15. *Berkenkamp F., Turchetta M., Schoellig A. P., Krause A.* Safe Model-based Reinforcement Learning with Stability Guarantees, 31st Conference on Neural Information Processing Systems, 2017. P. 908–919.

Отримано 28.10.2019,
в остаточному варіанті 12.11.2019