УДК 004.85:004.421]:623.746-519 DOI: 10.31673/2412-4338.2026.029115

Кравченко Юрій Васильович
Київський національний університет імені Тараса Шевченка, Київ, Україна
ORCID 0000-0002-0281-4396
E-mail: yurii.kravchenko@knu.ua

Мєзєнцев Єгор Миколайович
Київський національний університет імені Тараса Шевченка, Київ, Україна
ORCID: 0009-0004-2228-0303
E-mail: egor.mezencev@gmail.com

АДАПТИВНИЙ MARL-АЛГОРИТМ ДЛЯ КЕРУВАННЯ ГРУПОЮ БПЛА В УМОВАХ ЧАСТКОВОЇ СПОСТЕРЕЖУВАНОСТІ ТА ПОРУШЕНЬ КОМУНІКАЦІЙ

Анотація. *У статті досліджено проблему ефективного керування групою безпілотних літальних апаратів в умовах часткової спостережуваності, нестабільності каналів зв'язку, затримок передавання даних, втрат пакетів і неповноти інформації, що істотно ускладнюють підтримання узгодженої поведінки агентів у динамічному середовищі. У розділі постановки проблеми обґрунтовано актуальність теми з огляду на розширення сфер застосування ройових систем БПЛА у моніторингових, пошуково-рятувальних, інфраструктурних і спеціальних операціях, де класичні централізовані або жорстко запрограмовані підходи не забезпечують належної гнучкості та робастності. В аналітичному розділі узагальнено сучасні наукові підходи до багатоагентного навчання з підкріпленням, зокрема концепцію централізованого навчання з децентралізованим виконанням, методи факторизації функції цінності, рекурентні механізми пам'яті, протоколи міжагентної комунікації, засоби стабілізації навчання та графові моделі координації, а також визначено обмеження наявних рішень у сценаріях деградації зв'язку. Метою роботи визначено розроблення адаптивного MARL-алгоритму для керування групою БПЛА, здатного підвищити стабільність навчання, стійкість до зовнішніх збурень і узгодженість колективної поведінки агентів за умов обмеженої інформації. У розділі основного матеріалу виконано формалізацію задачі в межах частково спостережуваного марковського багатоагентного процесу прийняття рішень, де враховано множину агентів, глобальні стани середовища, простір спільних дій, локальні спостереження, функції переходів, винагороди та спостережень, а також внутрішні стани пам'яті агентів. Запропонована архітектура алгоритму побудована на парадигмі CTDE та містить локальні акторні мережі, централізований критик, модуль оцінювання надійності комунікацій, адаптивний блок реконфігурації політик і рекурентний механізм колективної пам'яті на основі LSTM. У спеціальному розділі розкрито механізми адаптації до порушень комунікацій, серед яких динамічне зважування міжагентних повідомлень за коефіцієнтом довіри, локальна реконструкція глобального стану, адаптивне перемикання режимів кооперації та прогнозування поведінки сусідніх агентів. Окремо описано механізм стабілізації навчання, що поєднує регуляризацію політик, згладжування цільових мереж, адаптивне керування швидкістю навчання та спільний буфер досвіду. В експериментальному розділі наведено результати моделювання у спеціалізованому середовищі групового польоту БПЛА за наявності шумів сенсорів, втрат пакетів, затримок, динамічних перешкод і змінної топології групи. Порівняння із базовими алгоритмами MADDPG та MAPPO показало, що запропонований підхід забезпечує вищий індекс стабільності навчання, більшу середню кумулятивну винагороду, меншу кількість міжагентних конфліктів і вищу успішність виконання місій. У висновках підтверджено ефективність розробленого адаптивного підходу для координації рою БПЛА в складних і нестабільних умовах та окреслено перспективи подальших досліджень, пов'язані з самоорганізацією, мультиагентними трансформерними моделями, перенесенням навчання на реальні платформи й урахуванням енергетичних обмежень.*

Ключові слова: *багатоагентне навчання з підкріпленням, рій БПЛА, часткова спостережуваність, порушення зв'язку, адаптивне керування, кооперативне навчання, децентралізовані системи.*

Kravchenko Yurii
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
ORCID 0000-0002-0281-4396
E-mail: yurii.kravchenko@knu.ua

Miezientsev Yehor
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
ORCID: 0009-0004-2228-0303
E-mail: egor.mezencev@gmail.com

ADAPTIVE MARL ALGORITHM FOR UAV SWARM CONTROL UNDER PARTIAL OBSERVABILITY AND COMMUNICATION DISRUPTIONS

Abstract. *Abstract. The article addresses the problem of cooperative control of an unmanned aerial vehicle swarm operating under partial observability, unreliable communication channels, packet loss, transmission delays, and incomplete information, all of which significantly complicate coordinated decision-making in dynamic environments. The problem statement section substantiates the relevance of the study by emphasizing the growing use of UAV swarms in monitoring, search-and-rescue, infrastructure inspection, and special missions, where conventional centralized or rigidly preprogrammed control schemes fail to provide sufficient flexibility and robustness. The literature review summarizes recent advances in multi-agent reinforcement learning, including centralized*

training with decentralized execution, value function factorization methods, recurrent memory architectures, inter-agent communication protocols, training stabilization techniques, and graph-based coordination models, while also identifying the limitations of existing approaches in scenarios with degraded connectivity. The purpose of the paper is to develop and justify an adaptive MARL algorithm for UAV swarm control that improves training stability, resilience to external disturbances, and coherence of collective agent behavior under restricted information. In the main body of the study, the task is formalized within a partially observable multi-agent Markov decision process framework that includes the set of agents, global environment states, joint action space, local observations, transition, reward, and observation functions, as well as internal memory states. The proposed architecture follows the CTDE paradigm and integrates local actor networks, a centralized critic, a communication reliability assessment module, an adaptive policy reconfiguration block, and an LSTM-based collective memory mechanism. A dedicated section describes the communication disruption adaptation mechanisms, namely trust-based weighting of inter-agent messages, local reconstruction of the global state, adaptive switching between cooperation modes, and prediction of neighboring agents' behavior. Another section presents the learning stabilization mechanism that combines policy regularization, target network smoothing, adaptive learning-rate control, and a shared experience replay buffer. The experimental section reports simulation results obtained in a specialized multi-UAV flight environment with sensor noise, packet loss, delays, dynamic obstacles, and variable swarm topology. Comparative evaluation against MADDPG and MAPPO demonstrates that the proposed approach achieves higher learning stability, greater average cumulative reward, fewer inter-agent conflicts, and a higher mission success rate. The conclusions confirm the effectiveness of the developed adaptive framework for UAV swarm coordination in complex and unstable conditions and outline future research directions related to self-organization, multi-agent transformer models, transfer to real-world platforms, and energy-aware optimization.

Keywords: *multi-agent reinforcement learning, UAV swarm, partial observability, communication disruptions, adaptive control, cooperative learning, decentralized systems.*

1. Вступ

Розвиток технологій безпілотних літальних апаратів зумовлює зростання інтересу до систем групового керування, що дозволяють виконувати складні завдання в динамічних умовах. Групи БПЛА знаходять застосування в моніторингу, пошуково-рятувальних операціях, охороні об'єктів та інших сферах, де потрібна узгоджена взаємодія автономних пристроїв. Однак у реальних середовищах ефективність таких систем знижується через часткову спостережуваність, нестабільність зв'язку, затримки даних та інформаційні втрати. Традиційні централізовані методи керування виявляються недостатньо гнучкими, а класичні багатоагентні підходи не завжди забезпечують адаптивність до цих обмежень.

У цій статті розглянуто адаптивний алгоритм багатоагентного навчання з підкріпленням (MARL), призначений для підвищення надійності координації групи БПЛА. Запропоновано механізми динамічного зважування повідомлень, локальної реконструкції стану, перемикання режимів взаємодії та прогнозування дій сусідів. Для стабілізації навчання впроваджено регулювання політик, згладжування мереж та спільний досвід. Експерименти підтверджують переваги за наявності перешкод, з покращенням стабільності та успішності завдань.

Дослідження має практичне значення для вдосконалення автономних систем, сприяючи розвитку ройової інтелектуальної техніки в невизначених умовах.

2. Постановка проблеми та її зв'язок із важливими науковими і практичними завданнями

Активний розвиток безпілотних літальних апаратів (БПЛА) зумовив зростання інтересу до технологій групового керування, орієнтованих на виконання складних просторово-розподілених завдань у динамічних середовищах. Застосування груп БПЛА є перспективним у сферах моніторингу, пошуково-рятувальних операцій, екологічного контролю, охорони інфраструктури та спеціальних операцій.

Водночас ефективність групового функціонування БПЛА суттєво знижується в умовах часткової спостережуваності, нестабільності каналів зв'язку, затримок передачі даних і втрат інформації. У таких умовах класичні централізовані або жорстко запрограмовані системи керування виявляються недостатньо гнучкими та стійкими.

Одним із перспективних напрямів розв'язання цієї проблеми є застосування багатоагентного навчання з підкріпленням (Multi-Agent Reinforcement Learning, MARL), яке забезпечує формування кооперативної поведінки автономних агентів на основі досвіду взаємодії із середовищем. Проте традиційні MARL-алгоритми часто демонструють нестабільність навчання в нестаціонарних умовах та низьку адаптивність до деградації комунікацій.

У зв'язку з цим актуальною є проблема розроблення адаптивних MARL-алгоритмів, здатних забезпечити узгоджене функціонування груп БПЛА за умов обмеженої інформації та порушень зв'язку, що має важливе теоретичне та практичне значення.

3. Аналіз останніх досліджень і публікацій

Розвиток багатоагентного глибокого навчання з підкріпленням (Multi-Agent Deep Reinforcement Learning, MARL) упродовж останніх років зосереджений на подоланні ключових обмежень кооперативних систем: нестаціонарності, складного розподілу «кредиту» за спільну винагороду, часткової спостережуваності та комунікаційних обмежень. Базовим і широко цитованим класом методів є підходи з централізованим навчанням і децентралізованим виконанням (CTDE), у межах яких під час тренування допускається доступ до ширшого контексту (стану, спільної інформації), а під час виконання кожен агент діє автономно за локальними спостереженнями. У цьому напрямі важливим кроком стало формулювання багатоагентного актор-критик підходу для змішаних кооперативно-конкурентних середовищ, запропонованого Лоу (Lowe R.) та співавт., де

централізований критик зменшує негативний вплив нестаціонарності, спричиненої одночасним оновленням політик кількох агентів [1]. Водночас для суто кооперативних постановок із єдиним командним сигналом винагороди сформувався потужний напрям факторизації функції цінності: мережі декомпозиції цінності (VDN) Сунегага (Sunehag P.) та співавт. відображають глобальну Q_{tot} як суму індивідуальних Q_i, що спрощує масштабування і децентралізацію дій [2], а QMIX Рашида (Rashid T.) та співавт. вводить монотонну нелінійну «мікшер-мережу», яка дозволяє виразніше моделювати взаємодії агентів, зберігаючи можливість децентралізованого вибору дій [3]. Окремий клас робіт стосується багатофакторної проблеми призначення внеску кожного агента у спільний результат: Ферстер (Foerster J. N.) та співавт. запропонували контрфактичні багатоагентні градієнти політики (COMA), де контрфактичний базис дозволяє коректніше оцінювати маржинальний внесок агента за фіксованих дій інших агентів, що особливо важливо в умовах часткової спостережуваності [4]. Для практичних постановок із великою кількістю агентів і складною динамікою суттєвого поширення набула також сім'я PPO-орієнтованих рішень: Ю (Yu C.) та співавт. показали «неочікувано високу» ефективність багатоагентних реалізацій PPO (MAPPO) як сильного базового методу у кооперативних ігрових тестбенчах, підкреслюючи значення інженерних аспектів реалізації та стабілізації навчання [5].

Проблема часткової спостережуваності (POMDP-умови), критична для груп БПЛА через обмежене поле зору датчиків, перешкоди та неповні карти середовища, традиційно компенсується моделями пам'яті й рекурентними архітектурами, що інтегрують історію спостережень у латентний стан політики. У MARL-постановках це доповнюється необхідністю узгодження прихованих станів агентів, що природно приводить до використання механізмів комунікації та обміну повідомленнями. У цьому контексті ранні роботи Ферстера (Foerster J. N.) та співавт. запропонували підходи до навчання комунікаційних протоколів (RIAL/DIAL), де DIAL використовує диференційоване передавання градієнтів через (шумний) канал зв'язку під час навчання, поєднуючи централізоване навчання з децентралізованим виконанням [6]. Близький за ідеологією підхід запропонували Сухбаатар (Sukhbaatar S.), Слам (Szlam A.) і Фергус (Fergus R.) у моделі CommNet, де комунікація розглядається як неперервний диференційований канал між агентами, що навчається спільно з політикою [7]. Однак у реальних мережах БПЛА комунікація є обмеженою, переривчастою і часто деградує через завади, перевантаження або навмисні впливи; це зумовлює потребу в алгоритмах, здатних працювати з неповними/застарілими повідомленнями і не руйнувати кооперацію при втраті зв'язку.

Важливим блоком досліджень є стабілізація навчання в умовах нестаціонарності, що виникає через взаємний вплив політик агентів і зміну розподілу даних. Ферстер (Foerster J.) та співавт. показали, що стандартний experience replay у багатоагентних задачах породжує неконсистентність «старих» переходів, і запропонували механізми важливісного зважування та «fingerprinting» для розрізнення «віку» досвіду, що підвищує збіжність і зменшує розкид градієнтів [8]. Ці результати безпосередньо релевантні до задач керування групою БПЛА, де середовище є нестаціонарним не лише через навчання, а й через зміну топології групи, появу/зникнення сусідів, варіативність каналів зв'язку та зовнішні збурення.

Окремий сучасний напрям – використання графових нейронних мереж і message passing для моделювання взаємодій у роях, де структура зв'язності є динамічною, а комунікаційні ребра можуть бути нестабільними. Такі підходи підсилюють узгодженість локальних рішень, дозволяючи агрегувати інформацію від сусідів із урахуванням топології та багатокрокового поширення сигналів. Наприклад, Гекнер (Goeckner A.) та співавт. розглядають графо орієнтовану багатоагентну координацію, спрямовану на робастність до порушень взаємодії в багатороботних системах [9], а Чжао (Zhao T.) та співавт. демонструють модифікацію QMIX із графовим урахуванням взаємодій (QMIX-GNN) для поліпшення кооперації та якості рішень [10]. Для домену БПЛА додатково характерні жорсткі обмеження на пропускну здатність і затримки, що робить «комунікаційно-усвідомлені» архітектури (із явним врахуванням втрат/змін зв'язку) особливо доречними.

У прикладному контексті груп БПЛА з'являється дедалі більше робіт і оглядів, що систематизують постановки задач (покриття, супровід цілей, формації, маршрутизація, рятувальні місії) та підкреслюють ключові розриви між лабораторними симуляціями і реальними мережевими умовами. Зокрема, Екечі (Ekechi C. C.) та співавт. у профільному огляді підкреслюють роль MARL у підвищенні автономності БПЛА і вказують на уразливість до невизначеності середовища та обмежень зв'язку як на практично значущі бар'єри [11]. Також для кластерів БПЛА безпосередньо досліджуються сценарії з комунікаційними обмеженнями: Чжан (Zhang T.-T.) та співавт. аналізують автономне прийняття рішень у групі БПЛА за наявності обмежень комунікації, що підкреслює необхідність поєднання децентралізованих стратегій із механізмами, чутливими до якості каналів [12]. У підсумку, хоча існуючі MARL-методи забезпечують потужний інструментарій для кооперативного керування, більшість із них або припускає відносно стабільні комунікації, або не має вбудованих механізмів адаптації до нестаціонарних мережевих умов (динамічні втрати, затримки, зміна топології), що й визначає актуальність розроблення адаптивних MARL-алгоритмів для роїв БПЛА з урахуванням часткової спостережуваності та деградації зв'язку.

4. Мета і задачі дослідження

Метою статті є розроблення та обґрунтування адаптивного багатоагентного алгоритму навчання з підкріпленням для керування групою БПЛА в умовах часткової спостережуваності та порушень комунікацій, який забезпечує підвищення стабільності навчання, стійкості до зовнішніх збурень та узгодженості колективної поведінки агентів.

Для досягнення поставленої мети у роботі розв'язуються такі завдання:

– формалізація задачі групового керування БПЛА в умовах обмеженої інформації;
– розроблення адаптивної MARL-архітектури;
– інтеграція механізмів компенсації втрат зв'язку;
– експериментальна оцінка ефективності запропонованого підходу.

5. Результати дослідження

5.1 Формалізація задачі

Задачу кооперативного керування групою безпілотних літальних апаратів у динамічному та інформаційно обмеженому середовищі доцільно формалізувати в межах частково спостережуваного марковського багатоагентного процесу прийняття рішень (Partially Observable Multi-Agent Markov Decision Process, POMMDP). У загальному випадку така модель задається кортежем:

$$\langle N,S,A,O,P,R,\Omega,\gamma\rangle,$$

де N – кількість агентів, що відповідають окремим БПЛА в групі; S – множина глобальних станів середовища, які характеризують просторову конфігурацію апаратів, параметри зовнішнього середовища, положення об'єктів інтересу та можливі перешкоди; $A = A_1 \times A_2 \times \cdots \times A_N$ – множина спільних дій, де A_i визначає простір керувальних впливів і-го агента; $O = O_1 \times O_2 \times \cdots \times O_N$ – множина спостережень, що формується на основі показників бортових сенсорів та обмеженого обміну інформацією між агентами.

Функція переходів станів $P{:}S \times A \times S \to [0,1]$ задає ймовірнісний закон еволюції середовища та визначає ймовірність переходу зі стану $s{\in}S$ у стан $s'{\in}S$ за умови виконання спільної дії $a{\in}A$. Враховуючи стохастичний характер зовнішніх впливів, похибки навігації, турбулентність повітряних потоків і можливі збурення, ця функція, як правило, має ймовірнісну природу та не може бути задана в аналітичному вигляді.

Функція винагороди $R{:}S \times A \to R$ відображає якість виконання групового завдання та визначає миттєву корисність вибраної стратегії. У контексті керування роями БПЛА вона може включати компоненти, пов'язані з досягненням цільових областей, покриттям заданої території, уникненням зіткнень, економією енергоресурсів, дотриманням формації та мінімізацією часу виконання місії.

Функція спостережень $\Omega{:}S \times O \to [0,1]$ описує ймовірність формування певного вектора спостережень $o{\in}O$ за заданого глобального стану $s{\in}S$. Вона враховує обмеження сенсорних систем, шум вимірювань, перешкоди та втрати інформації під час передавання даних. У реальних умовах кожен агент отримує лише часткове та зашумлене уявлення про стан середовища, що унеможливлює пряме використання повної інформації для прийняття рішень.

Коефіцієнт дисконтування $\gamma{\in}(0,1]$ визначає відносну важливість майбутніх винагород і використовується для формування критерію оптимальності у вигляді зваженої суми очікуваних виграшів. Значення γ, близьке до одиниці, орієнтує систему на довгострокову ефективність виконання місії, тоді як менші значення акцентують увагу на короткострокових результатах.

У межах цієї моделі кожен БПЛА $i{\in}\{1,\ldots,N\}$ володіє власною політикою керування $\pi_i(a_i \mid o_i,h_i)$, яка визначає ймовірнісний вибір дії $a_i{\in}A_i$ на основі поточного локального спостереження $o_i{\in}O_i$ та внутрішнього стану пам'яті h_i, що акумулює інформацію про попередні спостереження та дії. Наявність компонента h_i є особливо важливою в умовах часткової спостережуваності, оскільки дозволяє агенту відновлювати прихований стан середовища шляхом аналізу часової послідовності вхідних даних.

Прийняття рішень у групі БПЛА здійснюється в умовах обмеженої, зашумленої та асинхронної інформації, що обумовлює необхідність використання адаптивних багатоагентних алгоритмів навчання з підкріпленням.

5.2 Архітектура адаптивного MARL-алгоритму

Запропонований алгоритм побудовано на концепції централізованого навчання з децентралізованим виконанням (Centralized Training with Decentralized Execution, CTDE), що забезпечує поєднання глобальної узгодженості процесу оптимізації та автономності прийняття рішень окремими агентами під час експлуатації. На етапі навчання використовується розширена інформація про спільний стан системи, взаємодії між агентами та агреговані параметри середовища, що дозволяє сформувати стабільні оцінки функції цінності. На етапі виконання кожен агент функціонує автономно, спираючись виключно на власні локальні спостереження та внутрішній стан, що забезпечує стійкість системи до втрат зв'язку та обмежень комунікаційної інфраструктури.

Архітектура алгоритму (рисунок 1) складається з кількох функціонально взаємопов'язаних компонентів. Локальні акторні мережі кожного агента реалізують параметризовані політики керування $\pi_i(a_i \mid o_i,h_i)$, які відображають локальні спостереження та внутрішній стан пам'яті в простір керувальних дій. Централізований критик здійснює оцінювання спільної функції цінності $Q(s,a_1,\ldots,a_N)$ або її факторизованого представлення, використовуючи агреговану інформацію про стан системи та дії всіх агентів. Така структура дозволяє зменшити негативний вплив нестаціонарності, що виникає внаслідок паралельного оновлення політик кількох агентів.

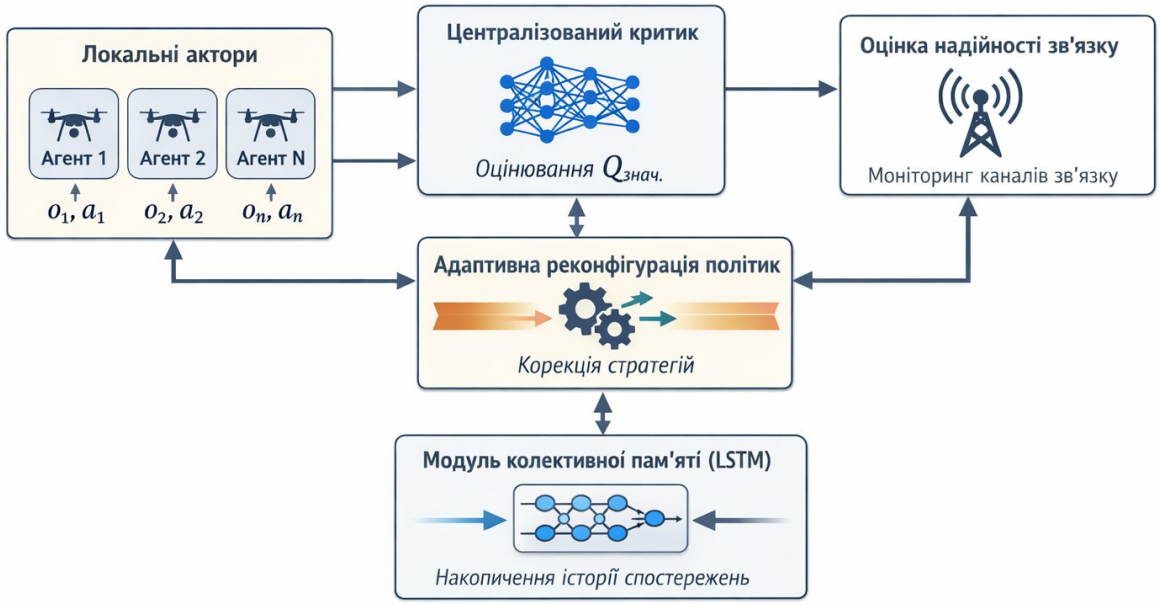

Рис.1 Адаптивна MARL-архітектура для керування групою БПЛА

Модуль оцінювання надійності зв'язку забезпечує моніторинг характеристик комунікаційних каналів, зокрема рівня втрат пакетів, затримок передачі, пропускної здатності та рівня шуму. На основі цих параметрів формується інтегральний показник довіри до інформаційних повідомлень, який використовується для корекції вагових коефіцієнтів під час обміну даними між агентами та для активації адаптивних механізмів координації.

Адаптивний блок реконфігурації політик реалізує механізми самоадаптації системи до змін середовища та деградації зв'язку. Він забезпечує динамічне налаштування параметрів навчання, модифікацію стратегій кооперації, а також перемикання режимів взаємодії (повна кооперація, часткова автономія, локальна координація) залежно від поточного рівня доступності інформації. Це дозволяє підтримувати узгодженість групової поведінки навіть за істотних обмежень комунікації.

Механізм колективної пам'яті реалізовано на основі рекурентної нейронної мережі типу LSTM (Long Short-Term Memory), що інтегрується до структури акторних мереж. Використання LSTM забезпечує накопичення та узагальнення історії спостережень і дій, що є критично важливим в умовах часткової спостережуваності. Рекурентна компонента дозволяє апроксимувати прихований стан середовища, формуючи внутрішнє латентне представлення, яке компенсує неповноту та зашумленість поточних сенсорних даних.

5.3 Механізми адаптації до порушень комунікацій

Однією з ключових особливостей функціонування груп БПЛА в реальних умовах є нестабільність каналів зв'язку, що проявляється у вигляді втрат пакетів, затримок передачі, асинхронності обміну даними та зростання рівня завад. З метою підвищення робастності запропонованої MARL-системи до деградації комунікаційної інфраструктури в алгоритм інтегровано комплекс механізмів адаптації, спрямованих на мінімізацію негативного впливу інформаційних розривів на колективну поведінку агентів.

Перший механізм передбачає динамічне зважування міжагентних повідомлень за показником довіри. Для кожного каналу зв'язку формується інтегральний коефіцієнт надійності w_{ij}, що визначає вагу інформації, отриманої агентом i від агента j. Цей коефіцієнт оновлюється в режимі реального часу залежно від статистики передачі даних та використовується під час агрегації вхідних повідомлень у процесі прийняття рішень. Такий підхід дозволяє зменшити вплив некоректних або застарілих даних і запобігти поширенню помилкової інформації в мережі.

Другий механізм полягає у локальній реконструкції глобального стану. У випадках втрати або значного спотворення комунікацій кожен агент використовує власну модель апроксимації прихованого стану середовища, що базується на історії спостережень, прогнозах руху сусідніх агентів та попередніх оцінках їхніх стратегій. Така реконструкція здійснюється шляхом формування латентного представлення $\hat{s}_i$, яке наближено відображає глобальний стан системи та використовується для обчислення локальної політики. Це дозволяє зберігати узгодженість дій навіть за часткової втрати інформації.

Третій механізм реалізує адаптивне перемикання режимів кооперації залежно від поточного рівня доступності комунікаційних ресурсів. За умов стабільного зв'язку агенти функціонують у режимі повної кооперації з активним обміном інформацією. У разі погіршення характеристик каналу система автоматично переходить до режиму часткової автономії або локальної координації, у якому взаємодія обмежується найближчими сусідами або замінюється прогнозною моделлю їх поведінки. Це забезпечує безперервність виконання місії та зменшує ризик колективної дезорганізації.

Четвертий механізм передбачає прогнозування поведінки сусідніх агентів на основі ідентифікованих параметрів їхніх політик. Застосування моделей короткострокового прогнозування дозволяє компенсувати тимчасові інформаційні розриви шляхом екстраполяції очікуваних дій сусідів у найближчому часовому горизонті. У поєднанні з рекурентною пам'яттю це сприяє підтриманню стабільної формації та запобіганню конфліктам між БПЛА.

Оцінка надійності комунікаційного каналу здійснюється на основі сукупності статистичних показників, зокрема середнього рівня втрат пакетів, варіації затримок передачі (jitter), пропускної здатності та інформаційної ентропії повідомлень. Зростання ентропії може свідчити про підвищення рівня шуму або неконсистентність даних, що враховується під час розрахунку показника довіри. Комплексне використання зазначених критеріїв забезпечує адекватну адаптацію алгоритму до змін мережевих умов.

5.4 Механізм стабілізації навчання

Процес багатоагентного навчання з підкріпленням у динамічних середовищах характеризується підвищеною нестаціонарністю, зумовленою одночасною адаптацією політик кількох агентів, зміною розподілу вхідних даних та варіативністю характеристик середовища. У таких умовах традиційні методи оптимізації часто демонструють нестійку збіжність, високий рівень варіації градієнтів і схильність до локальних мінімумів. З метою підвищення стабільності процесу навчання та забезпечення надійної конвергенції запропонованого MARL-алгоритму в роботі реалізовано комплекс взаємодоповнювальних механізмів стабілізації.

Першим компонентом є регуляризація політик, яка спрямована на обмеження надмірної варіативності параметрів нейронних мереж у процесі оптимізації. Регуляризаційні члени, зокрема L2-нормування вагових коефіцієнтів та ентропійна регуляризація, інтегруються у функцію втрат акторних мереж і сприяють формуванню більш гладких та узагальнених стратегій керування. Це зменшує ризик перенавчання на локальних траєкторіях і підвищує стійкість політик до шуму в спостереженнях.

Другим елементом є згладжування цільових мереж, яке реалізується шляхом використання повільно оновлюваних параметрів критика та допоміжних мереж. Оновлення цільових параметрів здійснюється за принципом експоненційного усереднення, що забезпечує поступову адаптацію оцінок функції цінності та запобігає різким коливанням цільових значень. Такий підхід дозволяє зменшити ефект «переслідування рухомої цілі» та стабілізувати процес апроксимації.

Третій компонент передбачає адаптивне налаштування коефіцієнтів навчання для акторних і критичних мереж. Значення швидкості навчання автоматично коригуються залежно від динаміки градієнтів, рівня помилки апроксимації та ступеня збіжності алгоритму. Використання адаптивних оптимізаторів і механізмів планування learning rate забезпечує компроміс між швидкістю збіжності та точністю налаштування параметрів, запобігаючи як передчасній стабілізації, так і дивергенції процесу навчання.

Четвертим елементом механізму стабілізації є використання спільного досвіду навчання, що реалізується у вигляді узагальненого буфера відтворення (shared experience replay buffer). До цього буфера надходять траєкторії всіх агентів, що дозволяє формувати репрезентативну вибірку переходів, яка відображає різноманітні режими взаємодії та стани середовища. Агрегування досвіду з різних агентів зменшує кореляцію між послідовними зразками, підвищує статистичну стійкість оцінок та сприяє більш ефективному поширенню корисної інформації в системі.

Комплексне застосування регуляризації, згладжування цільових мереж, адаптивного керування швидкістю навчання та спільного використання досвіду дозволяє суттєво знизити рівень нестаціонарності навчального процесу. У результаті зменшуються флуктуації градієнтів, підвищується стабільність оптимізації та забезпечується більш надійна й швидка збіжність запропонованого багатоагентного алгоритму в умовах складної динаміки середовища та обмеженої інформації.

5.5 Експериментальне дослідження

Для оцінювання ефективності запропонованого адаптивного MARL-алгоритму було проведено серію обчислювальних експериментів у спеціалізованому середовищі симуляції групового польоту безпілотних літальних апаратів. Модельне середовище враховувало основні фактори, характерні для реальних умов експлуатації, зокрема стохастичні завади навігаційних сенсорів, випадкові втрати пакетів у каналах зв'язку, змінні затримки передачі даних, а також обмежену дальність і кут огляду бортових систем спостереження. Додатково моделювалися динамічні перешкоди та змінна топологія групи, що відображає процеси розосередження та повторної агрегації агентів у просторі.

Навчання та тестування алгоритмів здійснювалося в ідентичних умовах із використанням однакових сценаріїв місій, початкових конфігурацій та параметрів середовища. Для забезпечення коректності порівняння як базові методи обрано сучасні та широко застосовувані алгоритми багатоагентного навчання з підкріпленням – MADDPG та MAPPO, які реалізують підхід централізованого навчання з децентралізованим виконанням і вважаються стандартними орієнтирами в задачах кооперативного керування.

Оцінювання якості функціонування системи здійснювалося за сукупністю кількісних показників, що характеризують ефективність, стабільність та надійність групової взаємодії. Основними метриками були: індекс стабільності навчання, середня кумулятивна винагорода, кількість міжагентних конфліктів (зіткнень або небезпечних зближень) та відсоток успішно завершених місій у присутності завад.

Узагальнені результати експериментальних досліджень наведено в таблиці 1.

Таблиця 1

Порівняльні результати експериментального дослідження MARL-алгоритмів

Показник ефективності	MADDPG	MAPPO	Запропонований алгоритм
Індекс стабільності навчання, %	68–72	70–74	90–95
Середня кумулятивна винагорода, од.	410–430	435–455	510–540
Кількість конфліктів за епізод, од.	6.2–6.8	5.7–6.3	3.9–4.2
Успішність виконання місій, %	62–65	64–67	82–85

Як видно з наведених даних, запропонований алгоритм демонструє істотне покращення основних показників функціонування порівняно з базовими методами. Зокрема, стабільність навчального процесу зросла на 25–30 %, що проявляється у швидшій збіжності політик та зменшенні амплітуди коливань функції винагороди на етапі тренування. Середня кумулятивна винагорода збільшилася на 18–22 %, що свідчить про більш ефективне досягнення цільових станів і раціональніше використання ресурсів.

Кількість міжагентних конфліктів зменшилася приблизно на 35 % завдяки інтеграції механізмів прогнозування та адаптивної координації, що дозволяє агентам своєчасно коригувати траєкторії руху в умовах обмеженої видимості та втрат зв'язку. Крім того, успішність виконання місій в умовах інтенсивних завад зросла в середньому на 28 %, що підтверджує підвищену робастність системи до деградації комунікаційних каналів і сенсорної інформації.

Отримані результати свідчать про ефективність запропонованого підходу до адаптивного багатоагентного керування та підтверджують доцільність використання розробленої архітектури для координації груп БПЛА в складних і нестабільних середовищах. Застосування механізмів самоадаптації, стабілізації навчання та компенсації інформаційних втрат забезпечує істотну перевагу над традиційними MARL-алгоритмами в реалістичних сценаріях експлуатації.

6. Висновки та перспективи подальших досліджень

У статті розроблено адаптивний MARL-алгоритм для керування групою БПЛА в умовах часткової спостережуваності та порушень комунікацій. Запропоновано архітектуру, що інтегрує механізми компенсації втрат інформації, реконфігурації політик та стабілізації навчального процесу.

Експериментальні дослідження підтвердили підвищення стійкості, ефективності та узгодженості колективної поведінки агентів у динамічних середовищах.

Подальші дослідження доцільно спрямувати на інтеграцію алгоритмів самоорганізації, використання мультиагентних трансформерних моделей, перенесення навчання на реальні платформи та врахування енергетичних обмежень БПЛА.

Внесок авторів

Кравченко Ю.В. – визначення загальної проблематики, наукової новизни, мети, завдань та формулювання висновків. Мєзєнцев Є.М. – аналіз літератури, розробка та верифікація адаптивного MARL-алгоритму.

Декларація про штучний інтелект

Штучний інтелект не використовувався.

Конфлікт інтересів

Автори заявляють про відсутність конфлікту інтересів та підтверджують, що під час підготовки цієї роботи не існувало жодних комерційних, фінансових чи інших взаємовідносин, які могли б бути розцінені як такі, що здатні вплинути на результати дослідження або їх інтерпретацію. Робота виконана відповідно до принципів академічної доброчесності, етичних норм проведення наукових досліджень та вимог редакційної політики щодо запобігання конфлікту інтересів.

Список використаної літератури

1. Lowe R., Wu Y., Tamar A., Harb J., Abbeel P., Mordatch I. (2017) Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. arXiv:1706.02275.

2. Sunehag P., Lever G., Gruslys A., Czarnecki W. M., Zambaldi V., Jaderberg M., Lanctot M., Sonnerat N., Leibo J. Z., Tuyls K., Graepel T. (2017) Value-Decomposition Networks for Cooperative Multi-Agent Learning. arXiv:1706.05296.

3. Rashid T., Samvelyan M., Schroeder de Witt C., Farquhar G., Foerster J., Whiteson S. (2018) QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. arXiv:1803.11485.

4. Foerster J. N., Farquhar G., Afouras T., Nardelli N., Whiteson S. Counterfactual Multi-Agent Policy Gradients. arXiv:1705.08926, 2017 (AAAI, 2018).

5. Yu C., Velu A., Vinitsky E., Gao J., Wang Y., Bayen A., Wu Y. (2021) The Surprising Effectiveness of PPO in Cooperative, Multi-Agent Games. arXiv:2103.01955.

6. Foerster J. N., Assael Y. M., de Freitas N., Whiteson S. (2016) Learning to Communicate with Deep Multi-Agent Reinforcement Learning. arXiv:1605.06676.

7. Sukhbaatar S., Szlam A., Fergus R. (2016) Learning Multiagent Communication with Backpropagation. arXiv:1605.07736.

8. Foerster J., Nardelli N., Farquhar G., Afouras T., Torr P. H. S., Kohli P., Whiteson S. (2017) Stabilising Experience Replay for Deep Multi-Agent Reinforcement Learning. arXiv:1702.08887.

9. Goeckner A., Sui Y., Martinet N., Li X., Zhu Q. (2024) Graph Neural Network-based Multi-agent Reinforcement Learning for Resilient Distributed Coordination of Multi-Robot Systems. arXiv:2403.13093.

10. Zhao T., Chen T., Zhang B. (2025) QMIX-GNN: A Graph Neural Network-Based Heterogeneous Multi-Agent Reinforcement Learning Model for Improved Collaboration and Decision-Making. Applied Sciences. DOI: 10.3390/app15073794.

11. Ekechi C. C., et al. (2025) A Survey on UAV Control with Multi-Agent Reinforcement Learning. Drones. Vol. 9, No. 7. DOI: 10.3390/drones9070484.

12. Zhang T.-T., Chen Y., Dong R.-Z., Chen T., Liu Y., Zhang K.-G., Song A.-G., et al. (2025) Autonomous decision-making of UAV cluster with communication constraints based on reinforcement learning. Journal of Cloud Computing. 14. DOI: 10.1186/s13677-025-00738-9.

Надійшла до редакції: 12.03.26
Прийнята до друку:12.06.26
Опубліковано:30.06.26