ЦИФРОВА ЕКОНОМІКА ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ
Digital Economy and Information Technologies

УДК 004.89:656.135

Сергій Островецький аспірант,
Кафедра комп'ютерних наук
Запорізький національний університет
Запоріжжя, Україна

БАГАТОФАКТОРНА МОДЕЛЬ ПОДОЛАННЯ ПРОБЛЕМИ РОЗРІДЖЕНИХ СИГНАЛІВ ДЛЯ ЗАДАЧІ МАРШРУТИЗАЦІЇ ТРАНСПОРТУ З ЧАСОВИМИ ВІКНАМИ

Анотація. Досліджено проблему розрідженої винагороди при навчанні Offline RL-агентів для задачі VRPTW. Аналіз 12617 ітерацій ALNS довів неефективність ізольованих бінарних сигналів, що генерують 97% неінформативних переходів. Запропоновано комплексну функцію Reward Shaping із врахуванням зміни вартості, кількості авто та логістичних штрафів. Абляційне дослідження підтвердило, що метод створює щільний градієнт (98% інформативних ітерацій), діючи як жорсткий контролер допустимості маршрутів.
Ключові слова: *VRPTW, комбінаторна оптимізація, навчання з підкріпленням, Offline RL, Reward Shaping, абляційне дослідження, ALNS*

I.ВСТУП

Формулювання проблеми дослідження у загальному вигляді та її зв’язок із важливими науковими та практичними завданнями. Сучасна логістика та ланцюги постачань функціонують в умовах постійного зростання обсягів перевезень та жорстких вимог до термінів доставки. Фундаментом для оптимізації цих процесів є задача маршрутизації транспорту з часовими вікнами (Vehicle Routing Problem with Time Windows, VRPTW). Її розв'язання дозволяє мінімізувати транспортні витрати та кількість задіяних автомобілів, одночасно гарантуючи обслуговування клієнтів у чітко задані інтервали часу. Через високу обчислювальну складність (NP-hard) для пошуку рішень на практиці найчастіше використовують метаевристичні підходи. Зокрема, алгоритм Adaptive Large Neighborhood Search (ALNS) [5], який ітеративно покращує маршрути за допомогою операторів руйнування та відновлення, довів свою дієвість. Проте традиційні механізми вибору операторів у ALNS мають реактивний характер і досягли межі своєї ефективності, оскільки не здатні проактивно адаптуватися до складних топологій графів.

Перспективним шляхом подолання цього бар'єру є інтеграція методів навчання з підкріпленням (Reinforcement Learning, RL) [6], що дозволяє автоматизувати пошук оптимальних стратегій вибору евристик. Особливо актуальним є підхід Offline RL, який відкриває шлях до навчання агентів на масивах попередньо зібраних історичних даних без потреби дороговартісної взаємодії із симулятором у реальному часі.

Однак процес навчання RL-агентів у задачах комбінаторної оптимізації наражається на серйозну перешкоду — проблему «розрідженої винагороди» (sparse reward), за якої

позитивний зворотний зв'язок генерується виключно при знаходженні нового глобального оптимуму. Хоча традиційно ця проблема вважається найбільш гострою для Online-режиму (де вона унеможливлює ефективне дослідження середовища агентом), в парадигмі Offline RL розрідженість сигналів створює концептуально іншу, але не менш критичну перешкоду. Оскільки навчання відбувається на фіксованому наборі статичних траєкторій, відсутність щільного зворотного зв'язку призводить до фундаментальної проблеми розподілу заслуг (credit assignment problem). Якщо винагорода зустрічається раз на сотні кроків, алгоритмам оновлення ваг вкрай складно визначити, яка саме проміжна маніпуляція з маршрутом насправді сприяла фінальному успіху. Це спричиняє затухання корисного сигналу, «інформаційний голод» нейронної мережі, проблему зникаючих градієнтів та унеможливлює ефективне навчання моделей на емпіричних даних.

Аналіз останніх досліджень і публікацій. Основоположні принципи метаевристики ALNS та архітектури операторів руйнування і відновлення були закладені в класичних працях S. Ropke та D. Pisinger [5]. Сучасні дослідження W. Kool [2] та M. Nazari [4] довели практичну можливість використання глибоких нейронних мереж (зокрема механізмів Attention) для розв'язання логістичних задач. Окремої уваги заслуговують ґрунтовні праці S. Levine [3], у яких формалізовано парадигму Offline RL. Цей підхід дозволяє перетворити навчання з підкріпленням на задачу Data-Driven оптимізації, що дозволяє використовувати масштабні набори емпіричних даних, таких як бази траєкторій пошуку [1].

Виділення невирішених раніше частин загальної проблеми. Попри значний розвиток теоретичної бази Offline RL [3], архітектура функції винагороди в задачах транспортної логістики залишається малодослідженою. Припускаючи, що загальна архітектура взаємодії агента з середовищем ALNS вже зведена до формату Марковського процесу прийняття рішень, критичним та невирішеним етапом залишається синтез ефективного зворотного зв'язку саме для статичних датасетів. Застосування наївних (ізольованих) бінарних нагород робить корисний сигнал занадто рідкісним для стабілізації багатошарових мереж. Досі не розв'язано проблему розробки математичної моделі комплексної «щільної» (dense) винагороди — інженерії функції винагороди (Reward Shaping), яка б об'єктивно оцінювала проміжні структурні зміни маршрутів (допомагаючи вирішити проблему credit assignment) та водночас жорстко контролювала дотримання логістичних бізнес-обмежень на кожному кроці пошуку.

Визначення мети та завдань дослідження. Метою статті є розробка та емпірична перевірка математичної моделі багатофакторної щільної функції винагороди (Reward Shaping) для подолання проблеми розріджених сигналів (зокрема проблеми розподілу заслуг) при навчанні Offline RL-агентів в алгоритмі ALNS для задачі VRPTW. Для досягнення поставленої мети визначено такі завдання:

1. розробити комплексну математичну модель формування щільної функції винагороди, що враховує проміжні зміни вартості, економію задіяного транспорту та жорсткі логістичні штрафи;
2. на основі відкритого масиву емпіричних даних (12 617 ітерацій) [1] провести абляційне дослідження (ablation study) для порівняльної оцінки розподілу навчальних сигналів при використанні базових розріджених функцій та запропонованої комплексної математичної моделі.

II.МЕТОДИ ТА МОДЕЛІ ДОСЛІДЖЕННЯ

Основним теоретичним інструментом дослідження виступає концепція інженерії функції винагороди (Reward Shaping) у глибокому навчанні з підкріпленням. Спираючись на існуючу парадигму, у якій процес пошуку алгоритму ALNS розглядається як марковський процес прийняття рішень, ключовим завданням є обчислення скалярного сигналу R_t,

який агент отримує на кожному кроці t за обрану дію (пару евристик руйнування та відновлення).

Емпіричною та практичною базою дослідження слугує відкритий масив даних «VRPTW-Search-Trajectories-Dataset» [1], що містить 12617 ітерацій марковських переходів алгоритму ALNS. Даний датасет акумулює повний спектр логістичних метрик на кожному кроці, що дозволяє проводити ретроспективний аналіз різних архітектур функцій винагороди без необхідності повторного запуску ресурсомістких симуляцій.

Для демонстрації проблеми розрідженості градієнтів було формалізовано три базові (наївні) моделі винагороди, які найчастіше зустрічаються в неадаптованих алгоритмах:

1. *сувора розріджена (Strict Sparse)*: $R_{sparse} = I_{best}$, де $I_{best} \in \{0, 1\}$ — індикатор оновлення глобального рекорду;
2. *локальна бінарна (Local Binary)*: $R_{imp} = I_{improvement}$, де $I_{improvement} \in \{0, 1\}$ — індикатор будь-якого зменшення вартості на поточному кроці;
3. *локальна неперервна (Local Continuous)*: $R_{cost} = \delta_{cost}$, де δ_{cost} — кількісна зміна загальної вартості маршруту.

Оскільки використання виключно ізольованих індикаторів не здатне забезпечити стабільне навчання агента з урахуванням логістичних обмежень, запропоновано комплексну багатофакторну математичну модель щільної функції винагороди R_{dense}. Модель агрегує топологічні, економічні та штрафні показники середовища за допомогою системи вагових коефіцієнтів:

$$R_{dense} = \omega_1 \cdot \delta_{cost} + \omega_2 \cdot I_{best} + \omega_3 \cdot \delta_{routes} + \omega_4 \cdot \delta_{len} - \omega_5 \cdot P_{tw} - \omega_6 \cdot P_{uns} \quad (1)$$

де δ_{cost} — зміна загальної вартості розв'язку (позитивне значення означає покращення);

I_{best} — бінарний індикатор знаходження нового глобального оптимуму;

δ_{routes} — зміна кількості задіяних транспортних засобів (вивільнення авто стимулюється потужним позитивним сигналом);

δ_{len} — зміна середньої довжини маршруту (стимулювання компактності топології);

P_{tw} — абсолютне значення накопиченого штрафу за порушення часових вікон (Time window feasibility penalty);

P_{uns} — відсоток необслугованих клієнтів у поточній конфігурації маршруту;

$\omega_1 \dots \omega_6$ — налаштовувані гіперпараметри, що визначають баланс між розвідкою, експлуатацією та дотриманням бізнес-обмежень.

Методологія оцінки ефективності розробленої моделі базується на проведенні абляційного дослідження (ablation study). За допомогою розроблених Python-скриптів масив із 12617 ітерацій [1] було пропущено через усі чотири моделі винагороди. Для усунення впливу мікро-шумів (float inaccuracies) запроваджено поріг чутливості $\epsilon = 10^{-3}$. Сигнал класифікувався як позитивний ($R > \epsilon$), негативний ($R < \epsilon$) або нейтральний ($|R| \leq \epsilon$). Основним критерієм ефективності обрано щільність інформативного сигналу — сумарну частку ітерацій, що генерують ненульовий градієнт для навчання нейронної мережі Offline RL-агента.

Для проведення симуляції та розрахунку функції R_{dense} у базовому програмному середовищі було зафіксовано набір вагових коефіцієнтів ($\omega_1 \dots \omega_6$). Ці параметри були підібрані емпіричним шляхом для забезпечення математичного балансу між заохоченням агента до оптимізації маршрутів та жорстким покаранням за порушення бізнес-обмежень задачі VRPTW.

Значення всіх гіперпараметрів, що використовувалися для генерації результатів абляційного дослідження, наведено у таблиці 1.

Таблиця 1. Гіперпараметри комплексної функції винагороди (Reward Shaping)

Параметр у коді	Позначення	Значення	Опис та вплив на сигнал
w_cost	ω_1	1,0	Базова вага за локальне покращення загальної вартості (відстані). Слугує одиничним мірилом (baseline).
w_best	ω_2	50,0	Високий бонус за оновлення глобального рекорду (is_new_best). Забезпечує потужний спайк винагороди при знаходженні SOTA-рішень.
w_routes	ω_3	100,0	Критичний бонус за вивільнення транспортного засобу. Найвищий пріоритет, оскільки вартість авто є найдорожчим ресурсом логістики.
w_avg_len	ω_4	0,5	Мікро-бонус за ущільнення маршрутів (зменшення avg_route_len). Допомагає боротися зі структурною фрагментацією.
w_feas_pen	ω_5	5,0	Штрафний мультиплікатор за кожну одиницю порушення часових вікон (feasibility_penalty).
w_unserved	ω_6	20	Жорсткий штраф за наявність клієнтів, які залишилися без обслуговування. Гарантує уникнення «жадібного» викидання складних вузлів.
epsilon	ϵ	10^{-3}	Поріг чутливості (0.001). Використовується для фільтрації обчислювального шуму (float inaccuracies). Сигнали, модуль яких менший за ϵ, класифікуються як нейтральні (R=0).

III.РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Для перевірки ефективності розробленої комплексної математичної моделі щільної функції винагороди (Reward Shaping) було проведено експериментальне моделювання розподілу навчальних сигналів. Аналіз здійснювався на масиві з 12 617 ітерацій марковських переходів з відкритого датасету «VRPTW-Search-Trajectories-Dataset». Оцінювання проводилося шляхом порівняння частоти генерування ненульового градієнта (позитивного чи негативного) для базових (розріджених) та запропонованої функцій.

Для забезпечення репрезентативності та усунення впливу мікро-шумів розрахунків сигнали, модулі яких не перевищують значення порогу чутливості ϵ, ідентифікуються як «нейтральні» (сліпа зона для градієнтного спуску). Результати абляційного дослідження представлено в таблиці 2.

Таблиця 2. Порівняльний аналіз розподілу сигналів винагороди (Ablation Study)

Функція винагороди (Baseline)	Позитивний сигнал (R > 0)	Негативний сигнал (R < 0)	Нейтральний сигнал (R = 0)
Sparse: is_new_best (тільки глобальні рекорди)	2,16%	0,00%	97,84%
Sparse: is_improvement (будь-яке локальне покращення)	15,36%	0,00%	84,64%
Continuous: cost_delta (неперервна зміна вартості)	15,36%	82,46%	2,18%
Dense: Reward Shaping (запропонована багатофакторна модель)	5,05%	92,80%	2,15%

Як свідчать отримані емпіричні дані, традиційна орієнтація алгоритмів навчання виключно на глобальні оптимуми (модель Sparse: is_new_best) робить простір станів критично неінформативним. У цьому сценарії 97,84% ітерацій є нейтральними «сліпими зонами», що робить процес оновлення ваг нейронної мережі на ранніх етапах практично неможливим через проблему зникаючих градієнтів.

Впровадження локальних бінарних індикаторів (Sparse: is_improvement) дозволяє збільшити частку ітерацій з позитивним підкріпленням до 15,36%, однак цей підхід повністю позбавлений негативного зворотного зв'язку (0,00%). Відсутність пенальті-сигналів стимулює агента до «жадібної» поведінки, не дозволяючи системі розпізнавати деструктивні топологічні патерни. Спроба використання виключно ізольованої функції зміни вартості (Continuous: cost_delta) суттєво підвищує інформативність простору (лише 2,18% нейтральних сигналів), проте така модель спрямовує агента оптимізувати лише кілометраж, ігноруючи при цьому критичні логістичні обмеження (кількість транспорту та часові вікна).

Запропонована багатофакторна математична модель Dense: Reward Shaping демонструє концептуально інший, асиметричний розподіл. Висока концентрація негативного сигналу (92,80%) зумовлена функцією жорсткого регуляризатора: алгоритм розпізнає та нещадно штрафує будь-які комбінації евристик, які призводять до порушення часових вікон або зростання кількості необслугованих клієнтів, навіть за умови локального скорочення довжини маршруту. Позитивний сигнал на рівні 5,05% діє як високоякісний «фільтр», який заохочує агента виключно до структурно валідних та комплексних покращень (зокрема, зменшення кількості задіяних автомобілів без порушення обмежень).

Такий підхід, що гарантує 97,85% інформативних (ненульових) ітерацій, формує стабільний та керований градієнт. Це створює оптимальні математичні умови для безперервного навчання Offline RL-агента, оскільки система здатна ефективно збалансувати процеси локального покращення маршруту та безумовного дотримання жорстких бізнес-обмежень задачі VRPTW.

IV.ВИСНОВКИ

Проведене дослідження теоретично обґрунтовує та емпірично доводить неефективність класичних розріджених (sparse) функцій винагороди при підготовці даних для навчання Offline RL-агентів у задачах маршрутизації транспорту (VRPTW).

Завдяки проведеному абляційному дослідженню на базі масиву з 12 617 марковських переходів алгоритму ALNS встановлено, що використання ізольованих індикаторів не здатне забезпечити нейронну мережу збалансованим градієнтним сигналом. Зокрема, орієнтація виключно на знаходження глобальних оптимумів залишає понад 97% простору станів неінформативним («сліпі зони»), що унеможливлює ефективне оновлення ваг моделі. З іншого боку, використання лише локальних метрик покращення вартості провокує агента ігнорувати жорсткі логістичні обмеження через відсутність зворотного пенальті-зв'язку.

Розроблена та валідована комплексна математична модель щільної функції винагороди (Reward Shaping) успішно розв'язує виявлену проблему інформаційного дефіциту. Інтеграція в єдину функціональну залежність економічних стимулів (скорочення вартості, економія транспортних засобів) та суворих штрафів (за порушення часових вікон і наявність необслугованих клієнтів) дозволила скоротити частку нейтральних сигналів до 2,15%.

Доведено, що запропонована багатофакторна функція діє як ефективний математичний фільтр: вона генерує масивний превентивний негативний сигнал (92,80%) для відсічення будь-яких комбінацій евристик, що руйнують допустимість (feasibility) розв'язку, та формує чітко каліброване позитивне підкріплення (5,05%) виключно за структурно валідні покращення топології маршрутів. Впровадження такої моделі винагороди створює необхідні передумови для стабільного, керованого та швидкого навчання проактивних нейромережевих розв'язувачів у сфері транспортної логістики.

V.ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

Перспективи подальших розвідок у зазначеному напрямку полягають у розробці алгоритмічних механізмів динамічного адаптування вагових коефіцієнтів (ω_i) запропонованої багатофакторної функції винагороди безпосередньо в процесі навчання (концепція Auto-Reward Shaping). Окрім цього, пріоритетним вектором є практичне дослідження впливу розробленої щільної функції (dense reward) на швидкість збіжності та стабільність градієнтів конкретних архітектур глибокого навчання (зокрема, Deep Q-Network, Proximal Policy Optimization та консервативного Q-навчання) під час розв'язання надвеликих інстанцій задачі VRPTW.

У якості постановки дискусійних питань науковій спільноті пропонується обговорення проблеми «штрафного бар'єру»: чи не призводять занадто жорсткі логістичні пенальті за порушення часових вікон до штучного звуження простору експлорації RL-агента на ранніх етапах навчання? Також відкритим та дискусійним залишається питання доцільності застосування методів оберненого навчання з підкріпленням (Inverse Reinforcement Learning) для автоматичного виведення ідеальної композитної функції корисності на основі аналізу «еталонних» траєкторій експертних розв'язувачів. Запрошуємо дослідників у галузі комбінаторної оптимізації та машинного навчання до відкритого обговорення окреслених теоретичних аспектів, а також до тестування та валідації розроблених моделей на базі існуючих відкритих масивів даних.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] С. В. Островецький, «VRPTW-Search-Trajectories-Dataset», GitHub, 2025. [Електронний ресурс]. Режим доступу: https://github.com/SerganO/VRPTW-Search-Trajectories-Dataset

[2] W. Kool, H. van Hoof, and M. Welling, «Attention, Learn to Solve Routing Problems!», in Proc. 7th Int. Conf. on Learning Representations (ICLR), New Orleans, LA, USA, 2019. [Online]. Available: https://openreview.net/forum?id=ByxBFsRqYm

[3] S. Levine, A. Kumar, G. Tucker, and J. Fu, «Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems», arXiv preprint arXiv:2005.01643, 2020.

[4] M. Nazari, A. Oroojlooy, L. V. Snyder, and M. Takáč, «Reinforcement learning for solving the vehicle routing problem» in Advances in Neural Information Processing Systems 31 (NeurIPS), 2018, 11p.

[5] S. Ropke and D. Pisinger, «An adaptive large neighborhood search heuristic for the pickup and delivery problem with time windows», Transportation Science, vol. 40, no. 4, pp. 455-472, 2006. DOI: 10.1287/trsc.1050.0135

[6] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

Отримано 25.12.2025 р.

MULTIFACTOR MODEL FOR OVERCOMING THE SPARSE REWARD PROBLEM IN THE VEHICLE ROUTING PROBLEM WITH TIME WINDOWS

Serhii Ostrovetskyi, student PHD
Department of Computer Science
Zaporizhzhia National University
Zaporizhzhia, Ukraine

Abstract. The sparse reward problem in training Offline RL agents for VRPTW is investigated. Analyzing 12617 ALNS iterations proved the inefficiency of isolated binary signals generating 97% non-informative transitions. A complex Reward Shaping function considering cost changes, fleet size, and penalties is proposed. The ablation study confirmed that the method creates a dense gradient (98% informative iterations), acting as a strict controller of route feasibility.

Keywords: *combinatorial optimization, reinforcement learning, Offline RL, Reward Shaping, ablation study, ALNS.*
.

Recieved 25.12.2025