Київський національний університет імені Тараса Шевченка
Міністерство освіти і науки України
Київський національний університет імені Тараса Шевченка
Міністерство освіти і науки України

Кваліфікаційна наукова
праця на правах рукопису

Джога Андрій Сергійович

УДК 519.2

Дисертація

Аналіз стратегій послідовного розподілу ресурсів у стохастичному середовищі

124 — Системний аналіз
Інформаційні технології

Подається на здобуття наукового ступеня
доктора філософії

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело.

_____ А. С. Джога

Науковий керівник: **Розора Ірина Василівна**,
доктор фіз.-мат. наук, доцент

Київ — 2023

АНОТАЦІЯ

Джога А. С. Аналіз стратегій послідовного розподілу ресурсів у стохастичному середовищі. — Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 124 — Системний аналіз. — Київський національний університет імені Тараса Шевченка; Міністерство освіти і науки України, Київський національний університет імені Тараса Шевченка, Київ, 2023.

Дисертаційне дослідження присвячене вивченню асимптотичних властивостей стратегій послідовного розподілу ресурсів у стохастичному середовищі зі спостереженнями, які мають бета-розподіл. Розглядається середовище, яке представлене моделлю стохастичного багаторукого бандита. Головною метою роботи є здійснення асимптотичного аналізу стратегій та розробка алгоритмів їх застосування. Також вивчається вплив додаткової інформації на ефективність стратегій.

Розглядається задача послідовного прийняття рішення в умовах невизначеності у контексті взаємодії між суб'єктом, що приймає рішення, так званим агентом, і зовнішнім середовищем. Ця взаємодія відбувається протягом скінченого горизонту. На кожному кроці агент обирає дію із заданої множини та у відповідь отримує винагороду від середовища. Метою агента є послідовний вибір таких дій, які призводять до найбільшої можливої сукупної винагороди за весь горизонт. Основним ускладненням в цій задачі є те, що параметри процесу винагороди не відомі заздалегідь.

У наведеному огляді літератури більшість результатів, опублікованих до цього часу робіт за даною темою, зосереджені на аналізі середовища

зі спостереженнями, які мають розподіл Бернуллі чи нормальний. Варто звернути увагу, що бета-розподіл є більш відповідним відображенням випадкової поведінки відсотків і пропорцій, тож може бути більш корисним у моделюванні клінічних випробувань, в системах маршрутизації мережі тощо. Отже, в нашій роботі розглядається стохастичне середовище зі спостереженнями, які мають бета-розподіл. Для аналізу оцінок ефективності стратегій використовуються оцінки хвостів субгауссових випадкових величин та їх властивості.

Були розглянуті стратегія на базі надійного інтервалу, баєсова та жадібна стратегії. Адаптовано алгоритми для цих стратегій та отримано асимптотичні оцінки очікуваних сукупних втрат для кожної з них. Наведено асимптотичний аналіз баєсової стратегії у стохастичному середовищі з додатковою інформацією та побудовано новий алгоритм з використанням зваження вибірки для зменшення шуму. Проведено математичне моделювання запропонованих у роботі алгоритмів стратегій у стохастичному середовищі з різними параметрами розподілів та кількістю дій. Одержані чисельні результати показали, що стратегії є асимптотично оптимальними згідно з отриманими оцінками верхніх границь втрат.

Результати проведеної роботи мають теоретичне значення та можливість практичного застосування. Отримані рішення дозволяють будувати оцінки ефективності стратегії у середовищі з бета-розподілом, яке представлене моделлю стохастичного багаторукого бандита. Подібні моделі мають великий потенціал використання в адаптивній маршрутизації для мінімізації затримок у мережі, динамічному ціноутворенні, клінічних випробуваннях тощо.

Ключові слова: послідовний розподіл ресурсів, теорія прийняття рішень, асимптотичний аналіз, задача багаторукого бандита, імовірнісні нерівності, баєсовий підхід, випадкові процеси.

SUMMARY

Dzhoha A. S. Sequential resource allocation in a stochastic environment. — Qualification scientific work in the form of manuscript.

Thesis for doctor of philosophy degree in speciality 124 — System analysis. — Taras Shevchenko National University of Kyiv; Ministry of Education and Science of Ukraine, Taras Shevchenko National University of Kyiv, Kyiv, 2023.

The focus of this dissertation study lies in the asymptotic analysis of policies within the context of sequential resource allocation tasks in a stochastic environment. This environment is characterized by a collection of beta distributions with unknown parameters, resembling a stochastic multi-armed bandit model. The central aim of this study is twofold: firstly, to conduct an asymptotic analysis of policies and adapt algorithms tailored to the specific environment; and secondly, to explore the impact of contextual information on the effectiveness of these strategies.

The problem at hand pertains to decision-making under uncertainty, involving a sequential interaction between a decision-maker (referred to as the agent) and the stochastic environment. This interaction unfolds over a finite time horizon. At each step, the agent selects an action from a predefined set and, in return, receives a reward from the environment. The agent's objective is to consistently select actions that maximize the cumulative reward over the entire horizon. The primary challenge in this task is the lack of prior knowledge regarding the parameters governing the reward process.

It's noteworthy that most existing literature on this topic has predominantly focused on analyzing environments with observations following Bernoulli or normal distributions. Nevertheless, the beta distribution offers a more suitable

representation for modeling random behavior related to percentages and proportions, making it especially relevant for applications such as clinical trial simulations and network routing systems. Therefore, our work delves into a stochastic environment where observations follow a beta distribution. To assess the efficiency of these strategies, we employ estimates of the tails of sub-Gaussian random variables and their associated properties.

The strategies under consideration encompass those based on confidence intervals, Bayesian methods, and greedy approaches. We adapt algorithms tailored to these strategies and derive asymptotic estimates of regrets. Mathematical modeling of these strategy algorithms is conducted within a stochastic environment featuring various distribution parameters and action counts. Our numerical results affirm that these strategies are asymptotically optimal, as substantiated by the upper bounds estimates.

The outcomes of this work hold both theoretical significance and practical applicability. The obtained estimates facilitate evaluations of strategy effectiveness in an environment characterized by a beta distribution, as exemplified by the stochastic multi-armed bandit model. Such models exhibit great potential in diverse applications, including adaptive routing to minimize network delays, dynamic pricing, and clinical trials, among others.

Key words: sequential resource allocation, decision-making theory, asymptotic analysis, multi-armed bandit problem, probabilistic inequalities, Bayesian inference, stochastic process.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Публікації, в яких опубліковано основні наукові результати дисертації

1. Dzhoha A. S. Multi-armed bandit problem under delayed feedback // Bulletin of Taras Shevchenko National University of Kyiv. Series: Physics and Mathematics. 2021. no. 1, P. 20–26.
2. Dzhoha A. S. Sequential resource allocation in a stochastic environment: an overview and numerical experiments // Bulletin of Taras Shevchenko National University of Kyiv. Series: Physics and Mathematics. 2021. no. 3, P. 13–25.
3. Dzhoha A. S, Rozora I. V. Multi-armed bandit problem with online clustering as side information // Journal of Computational and Applied Mathematics. 2023. Vol. 427, P. 115–132.
4. Dzhoha A. S., Rozora I. V. Beta upper confidence bound policy for the design of clinical trials // Austrian Journal of Statistics. 2023. Vol. 52, no. SI, P. 26–39.

Публікації, які засвідчують апробацію матеріалів дисертації

5. Джога А. С. Модель багаторукого бандита у стохастичному середовищі та чисельні експерименти // VIII Всеукраїнська науково-практична конференція «Інформаційні технології — 2021. Математичне моделювання та обчислювальні методи». Київ, Україна. 20 травня 2021. С. 176–177.

6. Dzhoha A. S., Lebedev E. O. Sequential resource allocation under multi-armed bandit model with delays // XXXVI International Conference «Problems of Decision Making under Uncertainties». Kyiv, Ukraine. May 11-14, 2021. P. 37–38.
7. Dzhoha A. S., Rozora I. V. Multi-armed bandit problem with online clustering as side information // International Congress of Computational Engineering and Sciences ESCO. Pilsen, Czech Republic. June 13-16, 2022.
8. Dzhoha A. S., Rozora I. V. Sequential resource allocation under multi-armed bandit model with online clustering as side information // Baltic-Nordic-Ukrainian Workshop on Survey Statistics. Tartu, Estonia. August 23-26, 2022. P. 42–43.
9. Dzhoha A. S., Rozora I. V. The upper confidence bound strategy for multi-armed bandit problem // XXXVII International Conference «Problems of Decision Making under Uncertainties». November 23-25, 2022. P. 42–43.
10. Dzhoha A. S., Rozora I. V. Multi-armed bandit policy under delays for the design of clinical trial // 6th Baltic-Nordic-Ukrainian Conference on Survey Statistics. Helsinki, Finland. August 21-25, 2023. P. 50–51.

ЗМІСТ

ВСТУП

Актуальність теми. Дисертаційне дослідження присвячене аналізу стратегій послідовного розподілу ресурсів у стаціонарному стохастичному середовищі зі спостереженнями, які мають бета-розподіл. Стохастичне середовище представлене моделлю багаторукого бандита зі скінченним горизонтом. Головну увагу приділено побудові алгоритмів стратегій для середовища, що досліджується. Наводиться асимптотичний аналіз ефективності стратегій.

У сучасному світі питання послідовного розподілу ресурсів вимагає ефективних методів та інструментів для отримання максимальної користі та оптимізації витрат. За останні 15 років кількість досліджень у цьому напрямку значно зросла, що пов'язано з розвитком цифрових технологій та їх стрімким поширенням на різні сфери нашого життя. Задача багаторукого бандита ([14]), що виникла в теорії ймовірностей і статистиці, є однією з ключових концепцій, яка має безпосереднє застосування в задачах послідовного розподілу ресурсів, де прийняття рішень відбувається в умовах невизначеності. Адаптивні стратегії послідовного розподілу ресурсів у середовищі, яке представлене моделлю багаторукого бандита, використовуються у багатьох сучасних системах, таких як адаптивні маршрутизації для мінімізації затримок у мережі ([9, 37, 61]), динамічне ціноутворення ([22, 58, 66]), системи рекомендацій ([21, 73, 57]), клінічні випробування ([20, 76, 2]) тощо.

Так, наприклад, у клінічних дослідженнях стандартною практикою тестування нових лікарських препаратів та лікувальних методик на цей час є рандомізоване контрольоване випробування, де пацієнтів випадковим чином розподіляють на дві або декілька груп з різними протоколами дій

(лікування) для визначення найкращого методу (ліків) чи підтвердження ефективності обраного. Такий підхід цілком задовольняє цілям наукових досліджень, але при цьому практично ігнорує інтереси окремого пацієнта, оскільки не всі учасники мають змогу отримати ефективне лікування під час проведення випробувань. Останнім часом набувають популярності дослідження, які розглядають адаптивні стратегії призначення лікування під час клінічних випробувань ([75, 77]), що дозволяє динамічно вивчати ефективність досліджуваних лікарських препаратів чи методів протягом усього випробування та відповідно коригувати поточний протокол лікування для кожної групи пацієнтів. Такі підходи дозволяють досягти адаптивності шляхом використання методів, розроблених на основі проблеми багаторукого бандита.

Більшість результатів, опублікованих до цього часу, здебільшого зосереджені на аналізі середовища зі спостереженнями, які мають розподіл Бернуллі чи нормальний. У той час бета-розподіл є більш відповідним відображенням випадкової поведінки відсотків і пропорцій, тож може бути більш корисним у моделюванні клінічних випробувань чи в системах маршрутизації мережі. Тому питання пошуку ефективних методів розв'язання задач багаторукого бандита залишається актуальним.

Ця дисертаційна робота досліджує стратегії послідовного розподілу ресурсів у стохастичному середовищі зі спостереженнями, які мають бета-розподіл, використовуючи оцінки хвостів субгауссових випадкових величин та їх властивості. Потенціал цього підходу демонструється в одній з опублікованих статей ([34]), де наведено симуляцію експерименту з використанням набору даних, отриманих у результаті реальних клінічних випробувань.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційну роботу виконано в рамках державних бюджетних дослідни-

цьких наукових тем № 19БП015-05 «Розробка алгоритмів і програмного забезпечення оптимізації сучасних систем зв'язку та систем керування запасами» (номер державної реєстрації 0119U100305) та 23БФ015-01 «Розробка стохастичних моделей, статистичних методів для аналізу та оптимізації систем у медичній та соціально-економічній сферах» (номер державної реєстрації 0123U101997) кафедри прикладної статистики факультету комп'ютерних наук та кібернетики Київського національного університету імені Тараса Шевченка.

Мета і завдання дослідження. Метою дисертаційної роботи є дослідження асимптотичних властивостей стратегій послідовного розподілу ресурсів у стаціонарному стохастичному середовищі зі спостереженнями, які мають бета-розподіл. Основними завданнями даної роботи є:

- асимптотичний аналіз ефективності стратегії на базі надійного інтервалу, баєсової та жадібної стратегій;
- побудова алгоритмів стратегій, які досліджуються, для середовища зі спостереженнями, які мають бета-розподіл;
- дослідження середовища з додатковою інформацією;
- моделювання оцінок отриманих з асимптотичного аналізу ефективності стратегій.

Об'єктом дослідження є стратегії послідовного розподілу ресурсів у стохастичному середовищі. *Предметом дослідження* є асимптотичні властивості цих об'єктів, тобто граничні властивості ефективності стратегій.

Методи дослідження. У роботі застосовуються методи теорії ймовірностей, математичної статистики, системного аналізу та теорії інформації. Для чисельного моделювання використовуються мова програмування Python [74] та статистичний пакет R [64].

Наукова новизна отриманих результатів. Наукова новизна цієї дисертаційної роботи полягає в розширенні та застосуванні відомих ре-

зультатів для асимптотичного аналізу стратегій послідовного розподілу ресурсів. Досліджується стохастичне середовище зі спостереженнями, які мають бета-розподіл. В аналізі використовуються оцінки хвостів субгаусових випадкових величин та їх властивості. Основними результатами є:

- адаптовано алгоритм для *стратегії на базі надійного інтервалу* та отримано асимптотичну оцінку очікуваних сукупних втрат;
- побудовано алгоритм для *баєсової стратегії*, отримано асимптотичну оцінку очікуваних сукупних втрат з залежністю від неоптимальності дій, отримано баєсову асимптотичну оцінку очікуваних сукупних втрат без залежності від неоптимальності дій та припущень щодо апріорного розподілу;
- отримано оцінку ефективності *жадібної стратегії* для випадку з двома діями та оптимальним вибором кількості досліджень простору варіантів;
- наведено асимптотичний аналіз *баєсової стратегії у стохастичному середовищі з додатковою інформацією* та побудовано новий алгоритм з використанням зважування вибірки для зменшення шуму;
- проведено моделювання розглянутих у роботі алгоритмів стратегій для підтвердження отриманих результатів;
- для математичного моделювання та побудови графіків було розроблено програмне забезпечення та бібліотеки, які опубліковані як ресурс з відкритим кодом [28, 29].

Практичне значення отриманих результатів. Отримані результати даної роботи дозволяють будувати оцінки ефективності стратегії у середовищі з бета-розподілом, яке представлене моделлю стохастичного багаторукого бандита. Подібні моделі мають широке застосування в адаптивній маршрутизації для мінімізації затримок у мережі, динамічному ціноутворенні, клінічних випробуваннях тощо.

Особистий внесок здобувача. Усі результати дисертаційної роботи одержані здобувачем самостійно. За результатами наукової праці здобувача було опубліковано чотири роботи у фахових виданнях [26, 27, 36, 34]. Дві роботи [36, 34] опубліковані у співавторстві з науковим керівником доцентом Розорою І. В., у яких Розорі І. В. належить загальне керівництво роботою.

Апробація матеріалів дисертації. Результати дослідження доповідалися та обговорювалися на наступних всеукраїнських та міжнародних конференціях:

1. VIII Всеукраїнська науково-практична конференція «Інформаційні технології — 2021. Математичне моделювання та обчислювальні методи», 20 травня 2021 року.
2. «XXXVI International Conference Problems of Decision Making under Uncertainties», 11-14 травня 2021 року, секційна доповідь.
3. «8th International Congress of Computational Engineering and Sciences ESCO», Пльзень, Чеська Республіка, 13-17 червня 2022 року, секційна доповідь.
4. «Baltic-Nordic-Ukrainian Workshop on Survey Statistics», Тарту, Естонія, 23-26 серпня 2022 року, секційна доповідь.
5. «XXXVII International Conference Problems of Decision Making under Uncertainties», 23-25 листопада 2022 року, секційна доповідь.
6. «6th Baltic-Nordic-Ukrainian Conference on Survey Statistics», Гельсінкі, Фінляндія, 21-25 серпня 2023 року, секційна доповідь.

Публікації. За результатами дослідження опубліковано

- 4 статті у періодичних фахових виданнях [26, 27, 36, 34], два з яких [36, 34] індексуються в наукометричних базах Scopus та Web of Science і входять до квартилів Q2 та Q4 відповідно, інші два [26, 27] — наукові фахові видання України;

- 6 тез доповідей на конференціях [35, 31, 32, 33, 30, 25].

Структура та обсяг дисертації. Дисертаційна робота складається зі вступу, сьоми основних розділів з підрозділами, висновків, списку використаних джерел (80 найменувань) і додатку, який містить список публікацій здобувача за темою дисертації та відомості про апробацію результатів. Повний обсяг дисертації становить 131 сторінок, основний текст займає 108 сторінок.

Зміст роботи. У **першому розділі** наведено огляд літератури та результати, отримані іншими авторами за розглянутою тематикою. Наводиться стислий огляд послідовного аналізу, а також місце в ньому послідовного розподілу ресурсів. Описується сучасний стан вивчення задач подібних до тих, що розглядаються в роботі.

У **другому розділі** надаються основні загальні означення та деякі додаткові твердження, які використано в дисертаційній роботі. Розглядається послідовний розподіл ресурсів у середовищі, яке представлене моделлю стохастичного багаторукого бандита. Моделюється послідовність прийняття рішення в умовах невизначеності у контексті взаємодії між суб'єктом, що приймає рішення за допомогою деякої стратегії, і зовнішнім середовищем. Ця взаємодія відбувається протягом T кроків, що є горизонтом. На кожному кроці t обирається дія I_t із заданої множини $\{1, \dots, N\}$, у відповідь середовище видає винагороду $\xi_t \in \{ x \in \mathbb{R} : x \geqslant 0 \}$. Метою стратегії є послідовний вибір таких дій, які призводять до найбільшої можливої сукупної винагороди $\sum_{t=1}^{T} \xi_t$. Модель стохастичного багаторукого бандита задається N-вимірним вектором $\boldsymbol{v} = (Q_1, \dots, Q_N)$, де для кожної дії Q_i позначає розподіл імовірностей з математичним сподіванням μ_i. Параметри розподілів не відомі заздалегідь. У цій моделі винагорода кожної дії є незалежною та однаково розподіленою. Оптимальна стратегія для будь-якої стохастичної моделі — це стратегія вибору дії з найвищою очікуваною

винагородою за весь горизонт.

У *підрозділі 2.1* наведено опис середовища зі спостереженнями, які мають Бернуллі- та бета-розподіл та сформульовані наступні означення.

Означення 2.1. *Клас середовища зі спостереженнями, які мають розподіл Бернуллі*, задається наступним чином:

$$\mathcal{V}^{\text{Bern}} = \{\, (\text{Bern}(p_i))_{i=1,\dots,N} : p \in [0,1]^N \,\},$$

де для всіх $i \in \{1,\dots,N\}$ випадкові величини $(\xi_t : t \in \{1,\dots,T\} \wedge I_t = i)$ мають розподіл Бернуллі з параметром p_i ($\xi_t \sim \text{Bern}(p_i)$), який задається розподілом

$$f_{\xi_t}(k; p_i) = \begin{cases} p_i & \text{якщо } k = 1, \\ 1 - p_i & \text{якщо } k = 0, \\ 0 & \text{інакше,} \end{cases}$$

для $k \in \{0,1\}$ з параметром $0 \leqslant p_i \leqslant 1$.

Означення 2.2. *Клас середовища зі спостереженнями, які мають бета-розподіл*, задається наступним чином:

$$\mathcal{V}^{\text{Beta}} = \{\, (\text{Beta}(\alpha_i, \beta_i))_{i=1,\dots,N} : \alpha, \beta \in \mathbb{R}^N \wedge \alpha_i > 0 \wedge \beta_i > 0 \,\},$$

де для всіх $i \in \{1,\dots,N\}$ випадкові величини $(\xi_t : t \in \{1,\dots,T\} \wedge I_t = i)$ мають бета-розподіл з параметрами α_i, β_i ($\xi_t \sim \text{Beta}(\alpha_i, \beta_i)$) та щільність розподілу ξ_t має вигляд:

$$f_{\xi_t}(x; \alpha_i, \beta_i) = \frac{1}{\text{B}(\alpha_i, \beta_i)} x^{\alpha_i - 1} (1 - x)^{\beta_i - 1}$$

для $x \in [0,1]$ з параметрами $\alpha_i > 0$, $\beta_i > 0$ і бета-функцією

$$\text{B}(\alpha_i, \beta_i) = \int_0^1 x^{\alpha_i - 1} (1 - x)^{\beta_i - 1}\, dx.$$

Означення 2.3. Позначимо множину дій через $\mathbb{M} = \{1,\dots,N\}$. Нехай для $t \in \{1,\dots,T\}$: $\Omega_t = (\mathbb{M} \times \mathbb{R})^t \subset \mathbb{R}^{2t}$, $\mathcal{F}_t = \mathfrak{B}(\Omega_t)$ та κ_t — це ймовірнісне

ядро ([62]) від $(\Omega_{t-1}, \mathcal{F}_{t-1})$ до $(\mathbb{M}, 2^{\mathbb{M}})$, що є функцією $\kappa_t : \Omega_{t-1} \times 2^{\mathbb{M}} \to [0, 1]$. Тоді *стратегія κ у стохастичному середовищі класу $\mathcal{V}$* є послідовністю $(\kappa_1, \kappa_2, \ldots, \kappa_T)$.

Означення 2.4. Маємо стратегію κ у стохастичному середовищі класу $\mathcal{V}$, яка взаємодіє з моделлю багаторукого бандита

$$\boldsymbol{v} = (Q_i : i \in \{1, \ldots, N\}) \in \mathcal{V},$$

де Q_i — це ймовірнісна міра на $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$. На кожному кроці $t \in \{1, \ldots, T\}$, маємо *послідовну взаємодію між агентом, який приймає рішення за допомогою стратегії κ, і зовнішнім стохастичним середовищем*, представленим моделлю $\boldsymbol{v}$ з наступними припущеннями:

- умовний розподіл винагороди ξ_t за умови, що маємо $I_1, \xi_1, I_2, \xi_2, \ldots, I_{t-1}, \xi_{t-1}, I_t$, є Q_{I_t} майже напевно;
- умовний розподіл імовірностей дії I_t за умови, що маємо $I_1, \xi_1, I_2, \xi_2, \ldots, I_{t-1}, \xi_{t-1}$, є таким розподілом майже напевно:

$$\kappa_t(\,\cdot \mid I_1, \xi_1, I_2, \xi_2, \ldots, I_{t-1}, \xi_{t-1}).$$

Розглядається параметрична модель багаторукого бандита, у якій розподіл Q_i залежить від деякого невідомого параметра θ_i, який набуває значення з множини Θ. Нехай $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N) \in \Theta^N$, тоді параметрична модель стохастичного багаторукого бандита задається наступним чином:

$$\boldsymbol{v} = \boldsymbol{v}_{\boldsymbol{\theta}} = \left(v_{\theta_1}, v_{\theta_2}, \ldots, v_{\theta_N}\right).$$

Параметрична модель стохастичного багаторукого бандита у класичному підході описується наступним чином:

- $\boldsymbol{\theta} \in \Theta^N$ розглядається як невідомий параметр;
- для всіх $i \in \{1, \ldots, N\}$ випадкові величини $(\xi_t : t \in \{1, \ldots, T\} \wedge I_t = i)$ — незалежні однаково розподілені, з розподілом імовірностей v_{θ_i} і математичним сподіванням μ_i;

- використовується позначення $\mathbb{P}_{\boldsymbol{\theta}}$ ($\mathbb{E}_{\boldsymbol{\theta}}$ відповідно).

Параметрична модель стохастичного багаторукого бандита з точки зору баєсового аналізу описується наступним чином:

- параметр $\boldsymbol{\theta}$ розглядається як випадкова величина з деяким апріорним розподілом Π_0 на Θ^N;
- для всіх $i \in \{1, \ldots, N\}$ за умови θ_i випадкові величини $(\xi_t : t \in \{1, \ldots, T\} \wedge I_t = i)$ — незалежні однаково розподілені, з розподілом імовірностей v_{θ_i} і математичним сподіванням μ_i;
- використовується позначення $\mathbb{P}_{\mathbf{\Pi_0}}$ ($\mathbb{E}_{\mathbf{\Pi_0}}$ відповідно).

За замовчуванням використовується класичний аналіз, $\mathbb{P}$ ($\mathbb{E}$ відповідно).

Підрозділ 2.2 присвячено асимптотичному аналізу втрат. Наведено означення втрат для стаціонарного стохастичного середовища. Отримано функцію очікуваних втрат з залежністю від неоптимальності дій для випадку середовища, що досліджується. Надана оцінка ефективності стратегії на основі рівномірного дослідження. Показані приклади неоптимальних стратегій та пошуку нижньої границі у найгіршому випадку.

Неоптимальністю дії i у стаціонарному стохастичному середовищі будемо називати наступний вираз:

$$\max_{j=1,\ldots,N} (\mu_j - \mu_i) .$$

Означення 2.5 ([54]). *У загальному випадку очікувані сукупні втрати* $\mathbb{E}[L]$ *при використанні стратегії* κ*, яка визначає послідовність вибору дій* $I_1^\kappa, \ldots, I_T^\kappa$ *на горизонті* T*, мають вигляд*

$$\mathbb{E}[L^\kappa(T)] = \mathbb{E}\left[\max_{i=1,\ldots,N} \sum_{t=1}^{T} \xi_{i,t} - \sum_{t=1}^{T} \xi_{I_t^\kappa,t}\right],$$

де $\xi_{i,t}$ *— це винагорода отримана від середовища на кроці* t *при виборі дії* i*.*

Означення 2.6. *У стаціонарному стохастичному середовищі очікувані сукупні втрати* $\mathbb{E}_{\boldsymbol{\theta}}[L]$ *за* T *кроків при використанні стратегії* κ *визначаються*

як

$$\mathbb{E}_{\boldsymbol{\theta}}\left[L^{\kappa}(T)\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[T \max_{i=1,\dots,N} \mu_i - \sum_{t=1}^{T} \xi_t\right] = T \max_{i=1,\dots,N} \mu_i - \mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{t=1}^{T} \xi_t\right].$$

Означення 2.7 ([55]). *З точки зору баєсового аналізу у стаціонарному стохастичному середовищі очікувані сукупні втрати* $\mathbb{E}_{\mathbf{\Pi_0}}[L]$ *за* T *кроків при використанні стратегії* κ *визначаються як*

$$\begin{aligned}\mathbb{E}_{\mathbf{\Pi_0}}\left[L^{\kappa}(T)\right] &= \mathbb{E}_{\mathbf{\Pi_0}}\left[T \max_{i=1,\dots,N} \mu_i - \sum_{t=1}^{T} \xi_t\right] = \\ &= \mathbb{E}_{\mathbf{\Pi_0}}\left[\mathbb{E}_{\mathbf{\Pi_0}}\left[T \max_{i=1,\dots,N} \mu_i - \sum_{t=1}^{T} \xi_t \mid \boldsymbol{\theta}\right]\right] = \\ &= \mathbb{E}_{\mathbf{\Pi_0}}\left[\mathbb{E}_{\boldsymbol{\theta}}\left[L^{\kappa}(T)\right]\right].\end{aligned}$$

Лема 2.1. *У середовищі, яке представлене моделлю стаціонарного стохастичного багаторукого бандита, зі скінченим горизонтом* T *і кількістю дій* N*, очікувані сукупні втрати* $\mathbb{E}[L]$ *з залежністю від неоптимальності дій при використанні стратегії* κ *визначаються як*

$$\mathbb{E}\left[L^{\kappa}(T)\right] = \sum_{i=1}^{N} \max_{j=1,\dots,N} (\mu_j - \mu_i)\, \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{I_t = i\}}\right].$$

Означення 2.9 ([54]). Стратегія κ є *рівномірно ефективною*, якщо її втрати задовольняють

$$\forall \boldsymbol{\theta} \in \Theta^N,\ \forall \alpha \in (0,1] : \lim_{T\to\infty} \frac{\mathbb{E}\left[L^{\kappa}(T)\right]}{T^{\alpha}} = 0.$$

Означення 2.10 ([54, 17]). Стратегія κ є *асимптотично оптимальною*, якщо

$$\limsup_{T\to\infty} \frac{\mathbb{E}\left[L^{\kappa}(T)\right]}{\log(T)} \leqslant \sum_{i=2}^{N} \frac{(\mu_1 - \mu_i)}{D_{\mathrm{KL}}(v_{\theta_i}, v_{\theta_1})},$$

де перша дія є оптимальною без втрати загальності та D_{KL} — розходження Кульбака-Ляйблера.

Теорема 2.1. *Розглядаються дві моделі багаторукого бандита v^+ і v^- з діями $(1/2, (1+C)/2)$ та $(1/2, (1-C)/2)$ відповідно, де $C > 0$ — деяка стала. У стохастичному середовищі маємо наступну нижню границю у найгіршому випадку для будь-якої стратегії:*

$$\max\left(\mathbb{E}\left[L_{v^-}(T)\right], \mathbb{E}\left[L_{v^+}(T)\right]\right) \geqslant \frac{\log(C^2T)}{16C}.$$

У *підрозділі 2.3* наведені імовірнісні нерівності для асимптотичного аналізу втрат та отримані додаткові властивості, які далі використовуються для дослідження ефективності стратегій.

Означення 2.12 ([16]). Центрована випадкова величина η є σ-субгауссовою, якщо для всіх $\lambda \in \mathbb{R}$ існує $\sigma > 0$, що має місце нерівність

$$\mathbb{E}\left[\exp(\lambda\eta)\right] \leqslant \exp\left(\frac{\lambda^2\sigma^2}{2}\right).$$

Наслідок 2.2. *Нехай η — центрована випадкова величина, яка має бета-розподіл. Тоді η є $1/2$-субгауссовою випадковою величиною.*

Теорема 2.2 ([16]). *Нехай η — σ-субгауссова випадкова величина. Тоді для всіх $\varepsilon \geqslant 0$ виконується наступна нерівність:*

$$\mathbb{P}(\eta \geqslant \varepsilon) \leqslant \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right).$$

Наслідок 2.3. *Розглянемо незалежні однаково розподілені випадкові величини $\eta_1, \eta_2, \ldots, \eta_n$, які мають бета-розподіл з математичним сподіванням μ. Тоді для всіх $\varepsilon \geqslant 0$ мають місце наступні нерівності:*

$$\mathbb{P}\left(\frac{1}{n}\sum_{j=1}^{n}\eta_j \geqslant \mu + \varepsilon\right) \leqslant \exp\left(-2n\varepsilon^2\right)$$

та

$$\mathbb{P}\left(\frac{1}{n}\sum_{j=1}^{n}\eta_j \leqslant \mu - \varepsilon\right) \leqslant \exp\left(-2n\varepsilon^2\right).$$

У **третьому розділі** наведено асимптотичний аналіз стратегії на базі надійного інтервалу у середовищі зі спостереженнями, які мають бета-розподіл. Адаптовано алгоритм та покращено оцінку ефективності стратегії для даного випадку. Проведено чисельні експерименти та наведені отримані дані.

Алгоритм 3.1. *Алгоритм стратегії на базі надійного інтервалу для середовища зі спостереженнями, які мають бета-розподіл.*

Крок 1. Покласти $t = 1$.

Крок 2. Якщо $t \leqslant N$, то покласти $I_t = t$ та перейти до кроку 5.

Крок 3. Для кожної дії $i \in \{1, \ldots, N\}$ покласти

$$U_i(t) = \frac{\sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = i\}} \xi_s}{\sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = i\}}} + \sqrt{\frac{\log(T)}{\sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = i\}}}}.$$

Крок 4. Призначити $I_t = \arg\max_{i=1,\ldots,N} U_i(t)$.

Крок 5. Виконати відбір ξ_t з розподілу, пов'язаного з дією I_t.

Крок 6. Якщо $t > T$, то закінчити виконання алгоритму. Інакше — збільшити t на 1 та перейти до кроку 2.

Теорема 3.2. *При використанні стратегії на базі надійного інтервалу за алгоритмом 3.1 має місце наступна нерівність:*

$$\mathbb{E}\left[L(T)\right] \leqslant 2 \sum_{i=2}^{N} (\mu_1 - \mu_i) + \frac{1}{2} \sum_{i=2}^{N} \frac{\log(T)}{\mu_1 - \mu_i}.$$

У **четвертому розділі** розглядається баєсова стратегія у середовищі зі спостереженнями, які мають бета-розподіл. Розроблено алгоритм для нашого випадку на базі існуючого алгоритму для середовища зі спостереженнями, які мають розподіл Бернуллі. Отримано оцінку ефективності стратегії з точки зору баєсового аналізу. Описано підхід з марковськими процесами прийняття рішень. Показані результати чисельних експериментів.

Алгоритм 4.1. *Алгоритм баєсової стратегії у загальному випадку.*

Крок 1. Покласти $t = 1$.

Крок 2. Якщо $t \leqslant N$, то покласти $I_t = t$ та перейти до кроку 5.

Крок 3. Зробити оцінку параметра $\hat{\boldsymbol{\theta}} = \mathcal{L}\left(\cdot \mid I_1, \xi_1, \ldots, I_{t-1}, \xi_{t-1}\right)$.

Крок 4. Призначити $I_t = \arg\max_{i=1,\ldots,N} \hat{\theta}_i$.

Крок 5. Виконати відбір ξ_t з розподілу, пов'язаного з дією I_t.

Крок 6. Якщо $t > T$, то закінчити виконання алгоритму. Інакше — збільшити t на 1 та перейти до кроку 2.

Наслідок 4.1. *Нехай використовується наступна ієрархічна модель для алгоритму баєсової стратегії в середовищі зі спостереженнями (ξ_t), які мають бета-розподіл:*

$$\eta_t | \xi_t \sim \mathit{Bern}(\xi_t), \quad t = 1, \ldots, T,$$
$$\xi_t \sim \mathit{Beta}(\alpha_{I_t}, \beta_{I_t}).$$

Тоді маємо наступну нерівність:

$$\mathbb{E}_{\boldsymbol{\theta}}\left[L(T)\right] \leqslant \left(\sum_{i=2}^{N} \frac{1}{(\mu_1 - \mu_i)^2}\right) \log(T).$$

Теорема 4.3. *Розглядається стохастичне середовище зі спостереженнями, які мають бета-розподіл. При використанні баєсової стратегії за алгоритмом 4.1 має місце наступна нерівність для баєсових втрат:*

$$\mathbb{E}_{\boldsymbol{\Pi_0}}\left[L(T)\right] \leqslant 8N + 4\sqrt{NT\log(T)}.$$

У **п'ятому розділі** розглядається жадібна стратегія у середовищі зі спостереженнями, які мають бета-розподіл. Наведено асимптотичний аналіз верхньої границі втрат. Отримано оцінку ефективності стратегії для випадку з двома діями та оптимальним вибором кількості досліджень простору варіантів. Отримано оцінку ефективності стратегії у загальному випадку. Наведені результати проведених чисельних експериментів.

Алгоритм 5.1. *Алгоритм жадібної стратегії з фіксованою кількістю C досліджень кожної дії.* Розглядається стохастичне середовище зі скінченим горизонтом T і кількістю дій N. Кожна дія $i \in \{1, \dots, N\}$ має деякий розподіл з невідомим математичним сподіванням μ_i. Вибираючи дію I_t, модель виконує відбір ξ_t з розподілу, пов'язаного з дією I_t та, як результат, реалізація вибірки стає доступною для стратегії.

Крок 1. Покласти C та $t = 1$.

Крок 2. Якщо $t \leqslant CN$, то покласти

$$I_t = (t \bmod N) + 1$$

та перейти до кроку 4.

Крок 3. Призначити

$$I_t = \arg\max_{i=1,\dots,N} \frac{\sum_{s=1}^{t} \mathbb{1}_{\{I_s = i\}} \xi_s}{\sum_{s=1}^{t} \mathbb{1}_{\{I_s = i\}}}.$$

Крок 4. Виконати відбір ξ_t з розподілу, пов'язаного з дією I_t.

Крок 5. Якщо $t > T$, то закінчити виконання алгоритму. Інакше — збільшити t на 1 та перейти до кроку 2.

Теорема 5.1. *Розглядається середовище з двома діями та неоптимальністю $\Delta\mu = |\mu_1 - \mu_2|$. Тоді при використанні жадібної стратегії має місце наступна нерівність:*

$$\mathbb{E}\left[L(T)\right] \leqslant C\Delta\mu + (T - 2C)\Delta\mu \exp\left(-C\left(\Delta\mu\right)^2\right).$$

Наслідок 5.1. *При оптимізації кількості досліджень простору варіантів жадібна стратегія має наступну оцінку втрат у середовищі з двома діями та неоптимальністю $\Delta\mu$:*

$$\mathbb{E}\left[L(T)\right] \leqslant \frac{1}{\Delta\mu} + \frac{1}{\Delta\mu} \log\left(T\left(\Delta\mu\right)^2\right).$$

Теорема 5.2. *Для стохастичного середовища з кількістю дій N при викори-*

станні жадібної стратегії має місце наступна нерівність:

$$\mathbb{E}\left[L(T)\right] \leqslant C\sum_{i=2}^{N}(\mu_1-\mu_i)+(T-CN)\sum_{i=2}^{N}(\mu_1-\mu_i)\exp\left(-C\left(\mu_1-\mu_i\right)^2\right).$$

Шостий розділ присвячений опису математичного моделювання, яке використовується для чисельних експериментів у попередніх розділах дисертації. Також наведено результати чисельних експериментів, у яких порівнюються розглянуті стратегії у середовищі зі спостереженнями, які мають бета-розподіл:

- стратегія на базі надійного інтервалу за алгоритмом 3.1;
- баєсова стратегія за алгоритмом 4.2;
- жадібна стратегія за алгоритмом 5.1.

У **сьомому розділі** розглядається середовище, у якому процес винагороди кожної дії залежить від деякої додаткової інформації. На прикладі моделі клінічного випробування з адаптивними стратегіями, де кожна дія представлена окремим препаратом, що досліджується, на результат може впливати певна інформація щодо суб'єкта випробування як, наприклад, вікова категорія чи цілий набір даних. Запропоновано нове формулювання проблеми, де додаткова інформація це інше спостереження та представлена результатом послідовного кластерного аналізу.

У *підрозділі 7.1* наведено асимптотичний аналіз баєсової стратегії у середовищі з додатковою інформацією та спостереженнями, які мають бета-розподіл. Отримано асимптотичну оцінку очікуваних сукупних втрат з залежністю від неоптимальності дій та у загальному випадку.

Теорема 7.1. *Розглядається стохастичне середовище зі скінченим горизонтом T, кількістю дій N та додатковою інформацією K. Кожна дія $i \in \{1,\dots,N\}$ має бета-розподіл з невідомим математичним сподіванням $\mu_{y,i}$ за умови інформації $y \in \{1,\dots,K\}$. Тоді при використанні баєсової стратегії*

за алгоритмом 7.1 має місце наступна нерівність:

$$\mathbb{E}\left[L(T)\right] \leqslant 14\sqrt{KNT}.$$

Теорема 7.2. *При використанні баєсової стратегії за алгоритмом 7.1 маємо наступну оцінку верхньої границі втрат з залежністю від неоптимальності дій:*

$$\mathbb{E}\left[L(T)\right] \leqslant \max_{y=1,\dots,K}\left(\sum_{i=2}^{N}\frac{1}{\left(\mu_{y,1}-\mu_{y,i}\right)^2}\right)^2 K\log\left(\frac{T}{K}\right).$$

У *підрозділі 7.2* розглянено вплив помилкової класифікації. Адаптовано алгоритм для баєсової стратегії.

Теорема 7.3. *У найгіршому випадку, коли маємо високу невпевненість в класифікації на кожному кроці, при використанні баєсової стратегії за алгоритмом 7.2 маємо наступну оцінку з залежністю від неоптимальності дій:*

$$\mathbb{E}\left[L(T)\right] \leqslant \frac{T}{N}\sum_{i=2}^{N}\max_{y=1,\dots,K}\left(\mu_{y,1}-\mu_{y,i}\right).$$

У **висновках** сформульовано основні результати дослідження.

Автор дисертації висловлює щиру вдячність своєму науковому керівнику — доценту Розорі Ірині Василівні — за постійну підтримку, важливі поради та безцінну допомогу в роботі; професору Лебєдєву Євгену Олександровичу — за віру та заохочення, допомогу з обранням напрямку дослідження та формулюванням перших задач; усьому колективу кафедри прикладної статистики за всебічну допомогу під час навчання. Окрему подяку автор висловлює своїй дружині за підтримку та надану можливість.

РОЗДІЛ 1
ОГЛЯД ЛІТЕРАТУРИ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1.1. Нарис історії вивчення питання послідовного розподілу ресурсів

Задача багаторукого бандита відноситься до області послідовного розподілу ресурсів, яка розглядається в теорії послідовного аналізу. Вивчення цієї проблеми належить до загальної концепції навчання з підкріпленням, що є однією з галузей у сфері штучного інтелекту.

Узагальнено задачу багаторукого бандита можна описати як пошук компромісу у виборі між дослідженням простору варіантів і використанням найоптимальнішого варіанту з раніше відомих для прийняття рішень у реальному часі в умовах невизначеності. Знаходження балансу між дослідженням та використанням є вкрай важливим для досягнення оптимального результату з найменшими втратами в довгостроковій перспективі. Адже система, яка завжди вибирає дослідження нових варіантів, нехтує можливістю отримання переваг від вже здобутих знань. З іншого боку, система, яка тільки використовує попередні знання, не в змозі адаптуватися до значних змін у зовнішньому середовищі для досягнення оптимального результату чи його покращення з часом. Ця проблема відома як дилема між дослідженням та використанням та до її розв'язку можна наблизитись за допомогою адаптивної стратегії послідовного розподілу ресурсів.

Як початок сучасної історії послідовного аналізу можна зазначити дослідження перевірки статистичних гіпотез, проведеного A. Wald [78], для знаходження методів отримання статистичних висновків з нефіксованим числом випробувань контролю якості продукції. Теоретичні дослідження

у послідовному аналізі та їх застосування знайшли відображення в роботах G. Barnard, F. Anscombe, C. Stein, J. Wolfowitz, W. Thompson, H. Robbins та інших.

Успіх теорії послідовного аналізу в області перевірки статистичних гіпотез дав поштовх дослідження послідовного оцінювання. J. Haldane [46] та C. Stein [70] описали, як деякі проблеми точкового та інтервального оцінювання можна розв'язувати за допомогою послідовного аналізу. У своїй роботі C. Stein продемонстрував, що використовуючи ці дослідження можливо отримати надійний інтервал з довжиною, яка задана експериментатором та не залежить від дисперсії розподілу.

Інша теорія, яка витікає з теорії послідовного аналізу та є фундаментальною для багатьох класів задач багаторукого бандита, це теорія оптимальної зупинки. Виокремлення цього напрямку започаткували роботи A. Wald & J. Wolfowitz [79] та K. Arrow, D. Blackwell & M. Girshick [6], присвячені дослідженню задач послідовної перевірки статистичних гіпотез. У цих роботах для розв'язку проблеми оптимальної зупинки був запропонований баєсовий підхід. Була представлена модель, де умовна особа, що приймає рішення, спостерігає послідовність $\{R_n, \mathcal{F}_n, n \geqslant 1\}$ при $\mathbb{E}\,|R_n| < \infty$ для усіх n, де n це горизонт, R_n — винагорода та $\mathcal{F}_n$ — фільтрація. На кожному кроці потрібно зробити вибір: зупинити відбір вибірки та отримати доступну винагороду R_n чи продовжити генерацію з метою отримати більшу винагороду в майбутньому. В цьому випадку оптимальне правило зупинки N можливо знайти через максимізацію очікуваної винагороди $\mathbb{E}[R_N]$. Для пошуку N можемо відштовхуватися від рівняння

$$V_n = \max\left(V_n, \mathbb{E}\left[V_{n+1} \,|\, \mathcal{F}_n\right]\right), n = 1, 2, \ldots,$$

де V — функція цінностей (також дивись підрозділ 4.4). Проблему оптимальної зупинки у загальному вигляді описав J. Snell [69], а R. Bellman [11] створив розділ математики, присвячений методам

розв'язування багатокрокових задач оптимального керування — динамічне програмування.

Подальші дискусії, присвячені послідовному розподілу ресурсів, були розвинуті у роботах W. Thompson [72] та H. Robbins [65], де автори аналізували задачу багаторукого бандита. Розглядався експеримент, де особа, що приймає рішення, на кожному кроці послідовно обирає одну з двох дій (дворукий бандит), та спостерігає послідовність винагород $\xi_1, \ldots, \xi_n$. З кожною дією пов'язаний розподіл, параметри якого не відомі заздалегідь. Головною метою експерименту було пошук стратегії, при якій послідовний вибір дій призводить до найбільшої можливої сукупної винагороди за n кроків $\sum_{t=1}^{n} \xi_t$. Результатом роботи стали опис математичної моделі задачі багаторукого бандита та формалізація поняття ефективності цих стратегій. Ці дослідження послужили початком вивчення задачі багаторукого бандита як окремого напрямку.

Згодом задачу багаторукого бандита почали досліджувати в контексті різних середовищ: стаціонарне стохастичне, де процеси винагороди кожної дії в моделі багаторукого бандита розглядаються як незалежні однаково розподілені випадкові величини; нестаціонарне, де параметри дій моделі змінюються з часом; марковське, де процес винагороди описаний ланцюгом Маркова; змагальне, де відсутні будь-які припущення щодо характеру процесів, пов'язаних з діями в моделі; та інші.

Один з важливих проривів в цій області був зроблений у роботі J. Gittins & D. Jones [44]. Автори цієї роботи використовували теорію динамічного програмування з геометричним знецінюванням для розв'язку задачі багаторукого бандита, яку описав H. Robbins. Результатом роботи став узагальнений розв'язок задачі багаторукого бандита за допомогою, так званих, індексів динамічного розподілу, що дозволило звести поставлену задачу до параметризованого сімейства розв'язків проблеми оптимальної зупинки. Недоліком цієї роботи є те, що запропонована стратегія не є оптимальною

на скінченному горизонті.

Більш детальний огляд розвитку послідовного аналізу та опис ранніх робіт, присвячених проблемі багаторукого бандита, можна знайти у роботах B. Ghosh та P. Sen [43], D. Siegmund [67].

У даній роботі найбільшу увагу приділяється стратегії послідовного розподілу ресурсів у середовищі, яке представлене моделлю стаціонарного стохастичного багаторукого бандита зі скінченним горизонтом. Додатково, в розділі 2 ми використовуємо зв'язок з моделлю змагального багаторукого бандита для пошуку нижньої границі. В розділі 4 наводиться опис марковської моделі. Розглянемо ці середовища більш детально у наступному підрозділі.

1.2. Класифікація та аналіз існуючих середовищ

Задача багаторуких бандитів — це послідовна взаємодія між суб'єктом, що приймає рішення, так званим агентом, та зовнішнім середовищем. Ця взаємодія відбувається протягом T кроків, де T — натуральне число, що називається горизонтом. На кожному кроці $t = 1, 2, \ldots, T$ агент обирає дію I_t із заданої множини $\{1, 2, \ldots, N\}$, у відповідь середовище видає винагороду $\xi_t \in \mathbb{R}_{\geqslant 0}$, де $\mathbb{R}_{\geqslant 0} \coloneqq \{\, x \in \mathbb{R} : x \geqslant 0 \,\}$. Вибір дії I_t залежить від історії попередніх виборів і їх результатів. Метою агента є послідовний вибір таких дій із заданої множини, які призводять до найбільшої можливої сукупної винагороди за T кроків.

Ключовим моментом у даній проблемі є те, що середовище не відоме для агента, тобто невідомий розподіл винагород кожної дії моделі. Все, що агенту відомо, це те, що справжнє середовище належить до певного класу середовищ.

Однією з метрик вимірювання ефективності стратегії агента є його втрати $L(T)$ за T кроків, що є різницею між очікуваною винагородою при

виборі оптимальної дії на кожному кроці та винагородою при виборі дій відповідно до прийнятої стратегії. Вперше ця метрика була запропонована у роботі T. Lai & H. Robbins [54]. Сукупні втрати є функцією від часу та можуть бути визначені як втрати знецінювання на нескінченному горизонті, або як сума втрат на скінченному горизонті. Чим швидше цільова стратегія у процесі використання наближається до оптимальної, тим повільніше зростають сукупні втрати. Оптимальна стратегія зводить до мінімуму сукупні втрати за будь-який часовий горизонт T.

За різними властивостями задачу багаторукого бандита можна поділити на багато категорій. За кількістю кроків розглядають моделі зі скінченним і нескінченним горизонтом. За кількістю дій проблему можна поділити на моделі з двома діями, N діями та нескінченною кількістю дій. З точки зору стаціонарності середовища виділяють моделі стаціонарні, де розподіл імовірностей винагород фіксований і незалежний, та нестаціонарні — розподіл імовірностей дій може змінюватись з часом. Також виокремлюють моделі, де процес винагороди кожної дії має залежність від додаткової інформації.

За характером процесу винагороди та враховуючи аналіз ефективності стратегій виокремлюють три фундаментальні постановки проблеми: стохастичну, змагальну і марковську.

Стохастична модель. У цій моделі винагорода кожної дії є незалежною та однаково розподіленою. Оптимальна стратегія для будь-якої стохастичної моделі — це стратегія вибору дії з найвищою очікуваною винагородою за весь горизонт.

Очікувані сукупні втрати $\mathbb{E}[L]$ при використанні стратегії κ, яка визначає послідовність вибору дій $I_1^\kappa, \ldots, I_T^\kappa$ на горизонті T, мають вигляд

$$\mathbb{E}\left[L^\kappa(T)\right] = \mathbb{E}\left[\max_{i=1,\ldots,N} \sum_{t=1}^{T} \xi_{i,t} - \sum_{t=1}^{T} \xi_{I_t^\kappa,t}\right],$$

де $\xi_{i,t}$ — це винагорода отримана від середовища на кроці t при виборі дії i.

Одною з перших запропонованих стратегій для розв'язання задачі багаторукого бандита у стохастичному середовищі була баєсова стратегія ([65]). У баєсових стратегіях прийняття рішення засновано на поточному апостеріорному розподілі параметрів дій. В роботах [49, 3] автори показали, що баєсова стратегія є асимптотично оптимальною для середовища зі спостереженнями, які мають розподіл Бернуллі. Оптимальність баєсової стратегії в середовищах зі спостереженнями, які мають розподіл з одним параметром, залежить від вибору апріорного розподілу ([51]).

В роботі [4] було показано, що у випадку середовища зі спостереженнями, які мають нормальний розподіл, баєсова стратегія не є оптимальною за мінімаксом (дивись наступний пункт зі змагальною моделлю).

У 2002 році в роботах [8, 59] автори представили стратегію на базі надійного інтервалу з використанням принципу «оптимізму в умовах невизначеності». Автори показали, що ця стратегія є асимптотично оптимальною для випадку, коли носій функцій розподілів, пов'язаних з діями, є обмеженим. Ця стратегія на кожному кроці t для всіх дій i обчислює значення індексу $U_i(t)$ на основі верхньої границі надійного інтервалу за нерівністю Чебишева, яке з великою ймовірністю є завищеною оцінкою невідомого математичного сподівання розподілу, пов'язаного з дією i. На кожному кроці вибирається дія з найбільшим значенням $U_i(t)$.

В роботах [41, 53] була запропонована модифікація, яка використовує нерівність Чернова [19] для побудови індексу (оцінки). Цей варіант стратегії є також асимптотично оптимальним, але потребує додаткового розв'язування задачі оптимізації під час роботи алгоритму.

Більшість результатів, опублікованих до цього часу, здебільшого зосереджені на аналізі середовища зі спостереженнями, які мають розподіл Бернуллі чи нормальний.

Змагальна модель. У такій моделі процес винагород не є випадковим. Послідовність винагород можна розглядати як вибрані умовним супротивником. За характером взаємодії супротивника виділяють два випадки: супротивник вибирає послідовність винагород на початку горизонту, тобто він не займається вивченням стратегії агента; супротивник обирає винагороди на кожному кроці, що часто формулюють в термінах теорії ігор, а також використовують критерій мінімаксу.

Як метрика вимірювання ефективності у змагальній моделі найчастіше використовуються втрати за найгіршим для обраної стратегії сценарієм. Нехай $\mathcal{P}$ — це множина усіх можливих послідовностей з множини $\{1, \dots, N\} \times \{1, \dots, T\}$ на $[0, 1]$ для усіх $T \in \mathbb{N}$, тоді очікувані втрати за найгіршим для обраної стратегії κ сценарієм на послідовності $P \in \mathcal{P}$:

$$\sup_{P \in \mathcal{P}} \mathbb{E}\left[L^{\kappa,P}(T)\right].$$

За таких умов найкраща стратегія та, яка досягає найменших втрат за найгіршим сценарієм.

Найменші очікувані втрати у найгіршому випадку, які може отримати будь-яка стратегія κ з усіх можливих $\mathcal{K}$, виражається через мінімакс втрати:

$$\inf_{\kappa \in \mathcal{K}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[L(T)^{\kappa,P}\right].$$

Основною стратегією для змагального багаторукого бандита є алгоритм на основі роботи P. Auer та інших [59]. Цей алгоритм призначає зважені ймовірності кожній дії на кожному кроці. Ці зважені ймовірності оновлюються відповідно до результатів попередніх виборів дій та втрат. Для оцінювання дій використовується критерій ваги у вигляді експоненціальної функції. Експонентне зростання допомагає значно збільшити вагу кращих дій. Ця та подібні стратегії ([40, 71]) не є оптимальними у випадку стохастичного середовища.

Марковська модель. У марковських моделях кожна дія асоційована з ланцюгом Маркова, а перехід до нового стану відбувається, коли цю дію вибирають. Ця модель була розглянута у роботі J. Gittins [45], де він довів, що найоптимальніша стратегія — це стратегія вибору найвищого індексу динамічного розподілу.

В іншому варіанті моделі марковських бандитів перехід дії у новий стан відбувається на кожному кроці незалежно від того, вибрана ця дія чи ні (англ. *restless markov bandit*). Ця модель вперше була представлена у роботі P. Whittle [80]. У даному випадку проблема не має загального розв'язку.

У стандартній марковській моделі дія $i = 1, 2, \ldots, N$ описується незвідним неперіодичним ланцюгом Маркова з дискретним часом, який приймає значення у скінченній множині станів S_i, r_s^i — винагорода у стані $s \in S_i$, $P_i = \{p_i(s, s'), s, s' \in S_i\}$ — матриця ймовірностей переходу дії i. Метою моделі є пошук послідовності дій, яка призводить до найбільшої можливої сукупної винагороди.

Знаходження оптимальної стратегії у марковській моделі може вимагати забагато обчислення у зв'язку з потенційно великим простором станів. Розв'язки для випадку дворукого бандита були наведені у роботах [39, 12].

На відміну від згаданих вище робіт далі ми будемо досліджувати стаціонарне стохастичне середовище зі спостереженнями, які мають бета-розподіл. У асимптотичному аналізі ефективності стратегій ми використовуємо оцінки хвостів субгауссових випадкових величин та їх властивості (дивись роботи В. Булдигіна та Ю. Козаченка [16], Ю. Козаченка, О. Погориляка, І. Розори та А. Тегзи [1]).

Описання загальних методів пошуку асимптотичних оцінок, які використовуються в даній дисертації, містяться у виданнях [14, 68], присвячених аналізу моделі багаторукого бандита. Отримані раніше результати аналізу стратегій, які ми розглядаємо, наведено на початку відповідних розділів.

РОЗДІЛ 2
ПОСЛІДОВНИЙ РОЗПОДІЛ РЕСУРСІВ У СТОХАСТИЧНОМУ СЕРЕДОВИЩІ

У даному розділі розглядаються стратегії послідовного розподілу ресурсів у середовищі, яке представлене моделлю стохастичного багаторукого бандита. Подається опис середовища зі стаціонарним стохастичним розподілом та розглядаються його основні властивості. Наводиться асимптотичний аналіз ефективності стратегій. Опис середовища та стратегій, які розглядаються у даному розділі, опубліковано у статті [27].

Розділ побудовано наступним чином. В підрозділі 2.1 наведено опис середовища зі спостереженнями, які мають Бернуллі- та бета-розподіл з необхідним мінімумом відомостей стосовно послідовного аналізу, та факти, що відіграють важливу роль у доведенні основних результатів розділу. Описується модель багаторукого бандита та її властивості у середовищі, що розглядаються. Надаються класи середовищ. Порівнюються баєсовий і класичний підходи до аналізу втрат та описуються відповідні параметричні моделі. Підрозділ 2.2 присвячений асимптотичному аналізу втрат. Наводиться функція втрат для стаціонарного стохастичного середовища. Отримано функцію очікуваних втрат із залежністю від неоптимальності дій. Надається оцінка ефективності стратегії на основі рівномірного розподілу. Показані приклади неоптимальних стратегій та пошуку нижньої границі у найгіршому випадку. В підрозділі 2.3 наведені імовірнісні нерівності для асимптотичного аналізу втрат та отримані додаткові властивості, які будуть використовуватись для дослідження ефективності стратегій. Надаються означення субгауссових випадкових величин та їх основні властивості. Отримані результати порівнюються з нерівністю Чебишева.

2.1. Математична модель багаторукого бандита у стохастичному середовищі

Нехай задано ймовірнісний простір $(\Omega, \mathcal{F}, \mathbb{P})$. Далі вважаємо, що усі випадкові величини розглядаються на ньому.

Розглядається послідовний розподіл ресурсів у середовищі, яке представлене моделлю стохастичного багаторукого бандита. Моделюється послідовність прийняття рішення в умовах невизначеності у контексті взаємодії між суб'єктом, що приймає рішення, так званим агентом, і зовнішнім середовищем. Ця взаємодія відбувається протягом T кроків, де T — натуральне число, що називається горизонтом. На кожному кроці $t = 1, 2, \ldots, T$ агент обирає дію I_t із заданої множини $\{1, 2, \ldots, N\}$, у відповідь середовище видає винагороду $\xi_t \in \mathbb{R}_{\geqslant 0}$. Вибір дії I_t залежить від історії попередніх виборів і їх результатів:

$$(I_1, \xi_1, I_2, \xi_2, \ldots, I_{t-1}, \xi_{t-1}).$$

Метою агента є послідовний вибір таких дій із заданої множини $\{1, 2, \ldots, N\}$, які призводять до найбільшої можливої сукупної винагороди за T кроків, тобто $\sum_{t=1}^{T} \xi_t$.

Модель стохастичного багаторукого бандита задається N-вимірним вектором $\boldsymbol{v} = (Q_1, Q_2, \ldots, Q_N)$, де N — це кількість можливих дій; кожна дія Q_i характеризується розподілом імовірностей з математичним сподіванням μ_i. При виборі дії I_t, модель виконує відбір ξ_t з розподілу, пов'язаного з дією I_t та, як результат, реалізація вибірки стає доступною для агента.

Таким чином, стратегія агента — це відображення з множини виборів та їх результатів в множину дій. Агент будує стратегію послідовної взаємодії з середовищем для збільшення сукупної винагороди. Відповідно, середовище — це відображення з послідовності виборів у винагороду. І агент, і середовище можуть приймати рішення (тобто обирати дії чи винагороди

відповідно) випадковим чином.

Основним ускладненням у проблемі багаторукого бандита є те, що модель $\boldsymbol{v}$ не відома заздалегідь. Агент може знати тільки клас середовища $\mathcal{V}$, до якого належить модель $\boldsymbol{v} \in \mathcal{V}$.

2.1.1. Опис класів середовищ і стратегій послідовного розподілу ресурсів

У цьому розділі ми розглядаємо класи середовищ зі стаціонарними стохастичними розподілами. Додамо означення даних класів у наступному вигляді:

$$\mathcal{V} = \{\, \boldsymbol{v} = (Q_i : i \in \{1, \dots, N\}) : Q_i \in \mathcal{Q}_i \, \forall i \in \{1, \dots, N\} \,\},$$

де ми припускаємо, що для кожної дії $i \in \{1, \dots, N\}$ існує множина розподілів $\mathcal{Q}_i$.

Означення 2.1. *Клас середовища зі спостереженнями, які мають розподіл Бернуллі*, задається наступним чином:

$$\mathcal{V}^{\text{Bern}} = \{\, (\text{Bern}(p_i))_{i=1,\dots,N} : p \in [0,1]^N \,\},$$

де для всіх $i \in \{1, \dots, N\}$ випадкові величини $(\xi_t : t \in \{1, \dots, T\} \wedge I_t = i)$ мають розподіл Бернуллі з параметром p_i ($\xi_t \sim \text{Bern}(p_i)$), який задається розподілом

$$f_{\xi_t}(k; p_i) = \begin{cases} p_i & \text{якщо } k = 1, \\ 1 - p_i & \text{якщо } k = 0, \\ 0 & \text{інакше,} \end{cases}$$

для $k \in \{0, 1\}$ з параметром $0 \leqslant p_i \leqslant 1$.

Означення 2.2. *Клас середовища зі спостереженнями, які мають бета-розподіл*, задається наступним чином:

$$\mathcal{V}^{\text{Beta}} = \{\, (\text{Beta}(\alpha_i, \beta_i))_{i=1,\dots,N} : \alpha, \beta \in \mathbb{R}^N \wedge \alpha_i > 0 \wedge \beta_i > 0 \,\},$$

де для всіх $i \in \{1, \ldots, N\}$ випадкові величини $(\xi_t : t \in \{1, \ldots, T\} \wedge I_t = i)$ мають бета-розподіл з параметрами α_i, β_i ($\xi_t \sim \text{Beta}(\alpha_i, \beta_i)$) та щільність розподілу ξ_t має вигляд:

$$f_{\xi_t}(x; \alpha_i, \beta_i) = \frac{1}{\text{B}(\alpha_i, \beta_i)} x^{\alpha_i - 1}(1 - x)^{\beta_i - 1}$$

для $x \in [0, 1]$ з параметрами $\alpha_i > 0$, $\beta_i > 0$ і бета-функцією

$$\text{B}(\alpha_i, \beta_i) = \int_0^1 x^{\alpha_i - 1}(1 - x)^{\beta_i - 1}\, dx.$$

Результатом взаємодії агента з середовищем є наступна скінченна сукупність елементів:

$$(I_1, \xi_1, I_2, \xi_2, \ldots, I_T, \xi_T).$$

Тобто, послідовність виборів дій $(I_t)_{t=1,\ldots,T}$ визначає стратегію агента. Введемо фільтрацію $\mathcal{F}_t \subset \mathcal{F}$ наступним чином:

$$\mathcal{F}_t = \sigma\left(I_1, \xi_1, I_2, \xi_2, \ldots, I_t, \xi_t\right), \tag{2.1}$$

де $\mathcal{F}_s \subset \mathcal{F}_t, \forall s < t, s, t \in \{1, \ldots, T\}$.

Для недетермінованої стратегії відбір I_t здійснюється з розподілу ймовірностей на системі дій $\{1, \ldots, N\}$, який є $\mathcal{F}_{t-1}$-вимірним. Додамо означення стратегії агента для нашого випадку з середовищем, яке представлене моделлю стаціонарного стохастичного багаторукого бандита.

Найменшою борелєвою σ-алгеброю на множені B будемо позначати $\mathfrak{B}(B)$.

Означення 2.3. Позначимо множину дій через $\mathbb{M} = \{1, \ldots, N\}$. Нехай для $t \in \{1, \ldots, T\}$: $\Omega_t = (\mathbb{M} \times \mathbb{R})^t \subset \mathbb{R}^{2t}$, $\mathcal{F}_t = \mathfrak{B}(\Omega_t)$ та κ_t — це ймовірнісне ядро ([62]) від $(\Omega_{t-1}, \mathcal{F}_{t-1})$ до $(\mathbb{M}, 2^{\mathbb{M}})$, що є функцією $\kappa_t : \Omega_{t-1} \times 2^{\mathbb{M}} \to [0, 1]$. Тоді *стратегія κ у стохастичному середовищі класу $\mathcal{V}$* є послідовністю $(\kappa_1, \kappa_2, \ldots, \kappa_T)$.

Означення 2.4. Маємо стратегію κ у стохастичному середовищі класу $\mathcal{V}$, яка взаємодіє з моделлю багаторукого бандита

$$\boldsymbol{v} = (Q_i : i \in \{1, \dots, N\}) \in \mathcal{V},$$

де Q_i — це ймовірнісна міра на $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$. На кожному кроці $t \in \{1, \dots, T\}$, маємо *послідовну взаємодію між агентом, який приймає рішення за допомогою стратегії κ, і зовнішнім стохастичним середовищем*, представленим моделлю $\boldsymbol{v}$ з наступними припущеннями:

- умовний розподіл винагороди ξ_t за умови, що маємо $I_1, \xi_1, I_2, \xi_2, \dots, I_{t-1}, \xi_{t-1}, I_t$, є Q_{I_t} майже напевно;
- умовний розподіл дії I_t за умови, що маємо $I_1, \xi_1, I_2, \xi_2, \dots, I_{t-1}, \xi_{t-1}$, є таким розподілом імовірностей майже напевно:
$$\kappa_t(\,\cdot \mid I_1, \xi_1, I_2, \xi_2, \dots, I_{t-1}, \xi_{t-1}).$$

2.1.2. Опис параметричної моделі багаторукого бандита

Ми розглядаємо параметричну модель багаторукого бандита, у якій розподіл Q_i залежить від деякого невідомого параметра θ_i, який набуває значення з множини Θ. Нехай

$$\boldsymbol{\theta} = (\theta_1, \dots, \theta_N) \in \Theta^N,$$

тоді параметрична модель стохастичного багаторукого бандита задається наступним чином:

$$\boldsymbol{v} = \boldsymbol{v}_{\boldsymbol{\theta}} = \left(v_{\theta_1}, v_{\theta_2}, \dots, v_{\theta_N}\right).$$

В даній роботі у більшості випадків ми використовуємо параметричну модель, де $\boldsymbol{\theta}$ розглядається як невідомий параметр за класичним підходом статистики. Додатково, у розділі 4 ми розглядаємо параметричну модель за баєсовим підходом, тобто розглядаємо параметр $\boldsymbol{\theta}$ як випадкову величину

з деякого апріорного розподілу Π. Наведемо різницю між цими двома підходами з точки зору моделі стохастичного багаторукого бандита та позначення, яке будемо використовувати далі.

Параметрична модель стохастичного багаторукого бандита у класичному підході описується наступним чином:

- $\boldsymbol{\theta} \in \Theta^N$ розглядається як невідомий параметр;
- для всіх $i \in \{1, \dots, N\}$ випадкові величини $(\xi_t : t \in \{1, \dots, T\} \wedge I_t = i)$ — незалежні однаково розподілені, з розподілом імовірностей υ_{θ_i} і математичним сподіванням μ_i;
- використовується позначення $\mathbb{P}_{\boldsymbol{\theta}}$ ($\mathbb{E}_{\boldsymbol{\theta}}$ відповідно).

Параметрична модель стохастичного багаторукого бандита з точки зору баєсового аналізу описується наступним чином:

- параметр $\boldsymbol{\theta}$ розглядається як випадкова величина з деяким апріорним розподілом Π_0 на Θ^N;
- для всіх $i \in \{1, \dots, N\}$ за умови θ_i випадкові величини $(\xi_t : t \in \{1, \dots, T\} \wedge I_t = i)$ — незалежні однаково розподілені, з розподілом імовірностей υ_{θ_i} і математичним сподіванням μ_i;
- використовується позначення $\mathbb{P}_{\mathbf{\Pi_0}}$ ($\mathbb{E}_{\mathbf{\Pi_0}}$ відповідно).

За замовчуванням будемо використовувати класичний аналіз, $\mathbb{P}$ ($\mathbb{E}$ відповідно).

2.2. Асимптотичний аналіз ефективності стратегій

Головною метою агента в представленій моделі є максимізація очікуваної сукупної винагороди $\mathbb{E}\left[\sum_{t=1}^{T} \xi_t\right]$, що еквівалентно мінімізації очікуваних сукупних втрат за T кроків.

Означення 2.5 ([54]). *У загальному випадку очікувані сукупні втрати* $\mathbb{E}[L]$ при використанні стратегії κ, яка визначає послідовність вибору дій $I_1^\kappa, \dots, I_T^\kappa$

на горизонті T, мають вигляд

$$\mathbb{E}\left[L^{\kappa}(T)\right] = \mathbb{E}\left[\max_{i=1,\dots,N}\sum_{t=1}^{T}\xi_{i,t} - \sum_{t=1}^{T}\xi_{I_t^{\kappa},t}\right],$$

де $\xi_{i,t}$ — це винагорода отримана від середовища на кроці t при виборі дії i.

Використовуючи фільтрацію $\mathcal{F}_t \subset \mathcal{F}$, автори [54] у загальному випадку виразили функцію втрат через залежність від неоптимальності дій з припущенням, що середовище має принаймні одну неоптимальну дію. Значення неоптимальності дій корисно у асимптотичному аналізу пошуку границь та буде введено для нашого випадку далі. Додамо означення втрат для середовища, яке представлене моделлю стаціонарного стохастичного багаторукого бандита.

Оптимальна стратегія для будь-якої стаціонарної стохастичної моделі — це стратегія вибору дії з найвищою очікуваною винагородою за весь горизонт.

Зауваження 2.1. Ми робимо припущення, що математичне сподівання існує і є скінченним для кожної дії $i \in \{1, \dots, N\}$ з мірою ймовірності Q_i та визначене як

$$\mu_i = \int_{-\infty}^{\infty} x\, dQ_i(x).$$

Означення 2.6. *У стаціонарному стохастичному середовищі очікувані сукупні втрати* $\mathbb{E}_{\boldsymbol{\theta}}\left[L\right]$ за T кроків при використанні стратегії κ визначаються як

$$\mathbb{E}_{\boldsymbol{\theta}}\left[L^{\kappa}(T)\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[T\max_{i=1,\dots,N}\mu_i - \sum_{t=1}^{T}\xi_t\right] = T\max_{i=1,\dots,N}\mu_i - \mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{t=1}^{T}\xi_t\right].$$

Означення 2.7 ([55])**.** *З точки зору баєсового аналізу у стаціонарному стохастичному середовищі очікувані сукупні втрати* $\mathbb{E}_{\boldsymbol{\Pi_0}}\left[L\right]$ за T кроків при

використанні стратегії κ визначаються як

$$\begin{aligned}\mathbb{E}_{\mathbf{\Pi_0}}\left[L^{\kappa}(T)\right] &= \mathbb{E}_{\mathbf{\Pi_0}}\left[T\max_{i=1,\dots,N}\mu_i - \sum_{t=1}^{T}\xi_t\right] = \\ &= \mathbb{E}_{\mathbf{\Pi_0}}\left[\mathbb{E}_{\mathbf{\Pi_0}}\left[T\max_{i=1,\dots,N}\mu_i - \sum_{t=1}^{T}\xi_t \mid \boldsymbol{\theta}\right]\right] = \\ &= \mathbb{E}_{\mathbf{\Pi_0}}\left[\mathbb{E}_{\boldsymbol{\theta}}\left[L^{\kappa}(T)\right]\right].\end{aligned}$$

Апостеріорним розподілом параметра $\boldsymbol{\theta}$ на кроці t є умовний розподіл імовірностей $\boldsymbol{\theta}$ за умови, що маємо історію попередніх виборів та їх результатів

$$I_1, \xi_1, I_2, \xi_2, \dots, I_{t-1}, \xi_{t-1}, I_t, \xi_t$$

і визначається як

$$\Pi_t\left(\boldsymbol{\theta}\right) = \mathcal{L}\left(\boldsymbol{\theta} \mid I_1, \xi_1, I_2, \xi_2, \dots, I_{t-1}, \xi_{t-1}, I_t, \xi_t\right).$$

Зауваження 2.2. У випадку, коли є потреба явно виразити середовище у визначенні втрат, будемо використовувати індекс: $L_{\mathcal{V}}$.

Неоптимальністю дії i у стаціонарному стохастичному середовищі будемо називати наступний вираз:

$$\max_{j=1,\dots,N}\left(\mu_j - \mu_i\right). \tag{2.2}$$

Тепер отримаємо очікувані сукупні втрати через залежність від неоптимальності дій, які будемо використовувати далі. Так як ми розглядаємо випадок зі стаціонарним стохастичним середовищим і скінченим горизонтом T та скінченою кількістю дій N, нам буде достатньо скористатися правилом повного математичного сподівання для суми неоптимальності дій.

Означення 2.8. Нехай B — випадкова подія. Її *індикаторною величиною* $\mathbb{1}_B$

є двозначна випадкова величина, яка позначається як

$$\mathbb{1}_B(\omega) = \begin{cases} 1 & \text{якщо } \omega \in B, \\ 0 & \text{якщо } \omega \notin B. \end{cases}$$

Лема 2.1. *У середовищі, яке представлене моделлю стаціонарного стохастичного багаторукого бандита, зі скінченим горизонтом T і кількістю дій N, очікувані сукупні втрати $\mathbb{E}[L]$ з залежністю від неоптимальності дій при використанні стратегії κ визначаються як*

$$\mathbb{E}\left[L^{\kappa}(T)\right] = \sum_{i=1}^{N} \max_{j=1,\dots,N} (\mu_j - \mu_i)\, \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{I_t = i\}}\right].$$

Доведення. Оскільки для будь-якого кроку $t \in \{1, \dots, T\}$ виконується

$$\sum_{i=1}^{N} \mathbb{1}_{\{I_t = i\}} = 1,$$

маємо наступний вираз очікуваних сукупних втрат з означення 2.6:

$$\begin{aligned} \mathbb{E}\left[L^{\kappa}(T)\right] &= T \max_{i=1,\dots,N} \mu_i - \mathbb{E}\left[\sum_{t=1}^{T} \xi_t\right] = \\ &= T \max_{i=1,\dots,N} \mu_i - \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i=1}^{N} \xi_t \mathbb{1}_{\{I_t = i\}}\right] = \\ &= \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i=1}^{N} \left(\max_{i=1,\dots,N} \mu_i - \xi_t\right) \mathbb{1}_{\{I_t = i\}}\right] = \\ &= \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{E}\left[\left(\max_{i=1,\dots,N} \mu_i - \xi_t\right) \mathbb{1}_{\{I_t = i\}}\right]. \end{aligned}$$

Використовуючи правило повного математичного сподівання та те, що математичне сподівання випадкової величини ξ_t на кроці t за умови I_t

дорівнює μ_{I_t}, з останньої рівності маємо

$$\begin{aligned}
\mathbb{E}\left[L^{\kappa}(T)\right] &= \sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[\left(\max_{i=1,\ldots,N}\mu_i - \xi_t\right)\mathbb{1}_{\{I_t=i\}}\right] = \\
&= \sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\left(\max_{i=1,\ldots,N}\mu_i - \xi_t\right)\mathbb{1}_{\{I_t=i\}} \mid I_t\right]\right] = \\
&= \sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{1}_{\{I_t=i\}}\,\mathbb{E}\left[\left(\max_{i=1,\ldots,N}\mu_i - \xi_t\right) \mid I_t\right]\right] = \\
&= \sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{1}_{\{I_t=i\}}\left(\max_{i=1,\ldots,N}\mu_i - \mu_{I_t}\right)\right] = \\
&= \sum_{i=1}^{N}\max_{j=1,\ldots,N}(\mu_j - \mu_i)\,\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}\right].
\end{aligned}$$

Що і треба було довести. □

Таким чином, ми отримали очікувані сукупні втрати з залежністю від неоптимальності дії $i \in \{1,\ldots,N\}$ помноженої на очікувану кількість виборів цієї дії i за весь часовий горизонт T.

За визначенням очікуваних сукупних втрат $\mathbb{E}\left[L^{\kappa}\right]$ виділимо їх наступні властивості для будь-якої стратегії κ у стаціонарному стохастичному середовищі:

- $\mathbb{E}\left[L^{\kappa}\right] \geqslant 0$, що випливає з означення втрат з залежністю від неоптимальності дій (лема 2.1), де

  $$\max_{j=1,\ldots,N}(\mu_j - \mu_i) \geqslant 0,$$
  $$\mathbb{E}\left[\mathbb{1}_{\{I_t=i\}}\right] \geqslant 0$$

  за визначенням;
- якщо виконується $\mathbb{E}\left[L^{\kappa}\right] = 0$, це свідчить, що агент знає оптимальну дію заздалегідь, тобто

  $$\mathbb{P}\left(\mu_{I_t} = \max_{i=1,\ldots,N}\mu_i\right) = 1$$

  для всіх $t = 1, 2, \ldots, T$;

- з означення 2.6 видно, що

$$\mathbb{E}\left[L^{\kappa}\right] \leqslant T \max_{i=1,\ldots,N} \mu_i.$$

Якщо ми припустимо, що винагорода приймає значення в границях замкненого проміжку $\xi_t \in [0,1]$, тоді маємо $\mathbb{E}\left[L^{\kappa}\right] \leqslant T$. Таким чином, наша головна мета це знаходження стратегій, для яких втрати як мінімум сублінійні на горизонті T для будь-якої моделі багаторукого бандита $\boldsymbol{v} \in \mathcal{V}$ заданого класу $\mathcal{V}$, тобто

$$\lim_{T\to\infty} \frac{\mathbb{E}\left[L^{\kappa}_{\mathcal{V}}(T)\right]}{T} = 0$$

як приклад формулювання мети для знаходження стратегії κ з сублінійними втратами. Існуючі асимптотичні оцінки розглянемо у наступному пункті.

Зауваження 2.3. Для випадку стаціонарної стохастичної моделі, коли випадкова величина $\xi_t \in [a,b]$ приймає значення не на відрізку $[0,1]$, ми можемо застосувати нормування:

$$\xi'_t = \frac{\xi_t - a}{b-a}.$$

2.2.1. Асимптотичні оцінки

У даній дисертації ми розглядаємо параметричні моделі з розподілами з експоненційного сімейства ([50]), тобто для розподілу ймовірності дії i, який характеризується одним параметром θ_i, щільність розподілу може бути представлена за допомогою деяких функцій $\eta(\theta)$, $T(x)$, $A(\theta)$, $h(x)$ у наступному вигляді:

$$f_{\xi}(x;\theta_i) = h(x)exp\left(\eta(\theta_i)\cdot T(x) + A(\theta_i)\right).$$

Це дозволяє використовувати розходження Кульбака-Ляйблера [52] (що також називають відносною ентропією) як оцінку того, наскільки одна

ймовірнісна міра Q_1 відрізняється від іншої Q_2:

$$D_{\mathrm{KL}}(Q_1, Q_2) = \begin{cases} \int \log\left(\frac{dQ_1}{dQ_2}(\omega)\right)\, dQ_1(\omega) & \text{якщо } Q_1 \ll Q_2, \\ \infty & \text{інакше,} \end{cases}$$

де $Q_1 \ll Q_2$ позначає абсолютну неперервність Q_1 за Q_2.

Зауваження 2.4. Розходження Кульбака-Ляйблера не відповідає вимогам статистичної метрики, бо є асиметричною мірою.

Приклад 2.1. Для розподілу Бернуллі, який входить в експоненційне сімейство, розходження Кульбака-Ляйблера має вигляд

$$D_{\mathrm{KL}}\big(\mathrm{Bern}(p_1), \mathrm{Bern}(p_2)\big) = p_1 \log\left(\frac{p_1}{p_2}\right) + (1-p_1)\log\left(\frac{1-p_1}{1-p_2}\right),$$

де p_1, p_2 — параметри першого та другого розподілів Бернуллі відповідно. Якщо використовувати значення $p_1 = 1/4$ та $p_2 = 1/2$, то легко побачити, що

$$D_{\mathrm{KL}}\big(\mathrm{Bern}(p_1), \mathrm{Bern}(p_2)\big) \neq D_{\mathrm{KL}}\big(\mathrm{Bern}(p_2), \mathrm{Bern}(p_1)\big).$$

Означення 2.9 ([54])**.** Стратегія κ є *рівномірно ефективною*, якщо її втрати задовольняють

$$\forall \boldsymbol{\theta} \in \Theta^N,\ \forall \alpha \in (0,1] : \lim_{T\to\infty} \frac{\mathbb{E}\left[L^{\kappa}(T)\right]}{T^{\alpha}} = 0.$$

Далі припустимо, що перша дія є оптимальною без втрати загальності, тобто $\arg\max_{i=1,\dots,N} \mu_i = 1$. Автори [54] показали, що для рівномірно ефективних стратегій кількість виборів кожної неоптимальної дії i має як мінімум логарифмічну складність:

$$\liminf_{T\to\infty} \frac{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{I_t = i\}}\right]}{\log(T)} \geqslant \frac{1}{D_{\mathrm{KL}}(v_{\theta_i}, v_{\theta_1})}. \tag{2.3}$$

Використовуючи очікувані сукупні втрати з залежністю від неоптимальності дії (лема 2.1), отримаємо

$$\liminf_{T\to\infty} \frac{\mathbb{E}\left[L^{\kappa}(T)\right]}{\log(T)} \geqslant \sum_{i=2}^{N} \frac{(\mu_1 - \mu_i)}{D_{\mathrm{KL}}(v_{\theta_i}, v_{\theta_1})}.$$

Таким чином маємо нижню границю для будь-якої стратегії.

Означення 2.10 ([54, 17]). Стратегія κ є *асимптотично оптимальною*, якщо

$$\limsup_{T\to\infty} \frac{\mathbb{E}\left[L^{\kappa}(T)\right]}{\log(T)} \leqslant \sum_{i=2}^{N} \frac{(\mu_1 - \mu_i)}{D_{\mathrm{KL}}(\upsilon_{\theta_i}, \upsilon_{\theta_1})},$$

де перша дія є оптимальною без втрати загальності.

У наступних розділах розглядаються асимптотично оптимальні стратегії у середовищі зі спостереженнями, які мають бета-розподіл.

2.2.2. Приклад неоптимальних стратегій

Для стохастичної моделі стратегія розв'язку задачі полягає у пошуку балансу між дослідженням простору варіантів і використанням оптимального варіанту з вже відомих для отримання найбільшої можливої сукупної винагороди за відведений горизонт. Якщо стратегія передбачає тільки дослідження, вона може бути ефективною лише у простих випадках.

Приклад 1. *Стратегія з рівномірним дослідженням дій* буде ефективною, коли усі дії з заданої множини $\{1, \ldots, N\}$ є оптимальними, тобто $\mu_1 = \mu_2 = \cdots = \mu_N$. Інакше втрати набувають лінійну складність:

$$\begin{aligned}\mathbb{E}\left[L(T)\right] &= \sum_{i=1}^{N} \max_{j=1,\ldots,N} (\mu_j - \mu_i)\, \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{I_t = i\}}\right] = \\ &= \frac{T}{N} \sum_{i=1}^{N} \max_{j=1,\ldots,N} (\mu_j - \mu_i)\,.\end{aligned}$$

Приклад 2. *Стратегія, яка використовує тільки існуючі знання*, нехтуючи дослідженням можливих змін у середовищі може зазнати значних втрат. Наприклад, стратегія моделі стохастичного дворукого бандита з двома діями ($N = 2$), яка за перші два кроки вибирає першу та другу дії відповідно, а на всіх інших кроках $t \in \{3, 4, \ldots, T\}$ використовує найкращу

дію за вибірковим середнім,

$$I_t = \begin{cases} t & \text{якщо } t \leqslant N, \\ \arg\max_{i=1,\dots,N} \frac{\sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = i\}} \xi_s}{\sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = i\}}} & \text{інакше,} \end{cases}$$

може отримати лінійні втрати. Для прикладу візьмемо розподіли Бернуллі з параметрами $p_1 = 1/2$ та $p_2 < 1/2$ для першої та другої дій відповідно. Тоді з імовірністю $p_1 p_2$, стратегія отримає винагороди 0 та 1 за першу та другу дії відповідно, і буде змушена обирати неоптимальну другу дію на всіх інших кроках $t \in \{3, \dots, T\}$. Звідси втрати набувають наступного вигляду:

$$\mathbb{E}\left[L(T)\right] \geqslant p_1 p_2 \left(\frac{1}{2} - p_2\right)(T-1).$$

На цих прикладах видно, що для досягнення балансу між дослідженням і використанням потрібно використовувати відведений горизонт більш ефективно.

2.2.3. Приклад пошуку нижньої границі втрат

В продовження попереднього прикладу аналізу стратегій у середовищі зі спостереженнями, які мають розподіл Бернуллі, з двома діями, розглянемо моделі багаторукого бандита з параметрами дій $p_1 = 1/2$ та $p_2 = (1 \pm C)/2$ для пошуку нижньої границі у найгіршому випадку при використанні будь-якої стратегії, де $C > 0$ — деяка стала.

Будемо використовувати метод пошуку нижньої границі з роботи [42], де у 2017 році автори показали логарифмічну складність алгоритмів у середовищі зі спостереженнями, які мають розподіл Бернуллі у загальному випадку з $N \geqslant 2$, та застосуванням моделі змагальних багаторуких бандитів. Було запропоновано використання нерівності Bretagnolle-Huber. Через те, що у нашому випадку тільки дві дії, аналіз є значно простішим та

ґрунтується тільки на розходженні Кульбака-Ляйблера в додаток до використання нерівності Bretagnolle-Huber.

Означення 2.11. Нехай $\mathcal{K}$ — множина усіх стратегій. *Очікувані сукупні втрати у найгіршому випадку* у середовищі класу $\mathcal{V}$ визначаються за допомогою мінімаксу наступним чином:

$$\inf_{\kappa\in\mathcal{K}}\sup_{\boldsymbol{v}\in\mathcal{V}}\mathbb{E}\left[L_{\boldsymbol{v}}^{\kappa}(T)\right].$$

Лема 2.2 (Нерівність Bretagnolle-Huber, альтернатива нерівності Пінскера, [13])**.** *Маємо наступну нерівність для деяких ймовірнісних мір Q_1 і Q_2 на вимірному просторі $(\Omega,\mathcal{F})$ для події $B\in\mathcal{F}$:*

$$Q_1(B)+Q_2(B^c)\geqslant\frac{1}{2}\exp\left(-D_{\mathrm{KL}}(Q_1,Q_2)\right),$$

де D_{KL} — розходження Кульбака-Ляйблера та $B^c=\Omega\setminus B$.

Теорема 2.1. *Розглядаються дві моделі багаторукого бандита v^+ і v^- з діями $(1/2,(1+C)/2)$ та $(1/2,(1-C)/2)$ відповідно, де $C>0$ — деяка стала. У стохастичному середовищі маємо наступну нижню границю у найгіршому випадку для будь-якої стратегії:*

$$\max\left(\mathbb{E}\left[L_{v^-}(T)\right],\mathbb{E}\left[L_{v^+}(T)\right]\right)\geqslant\frac{\log(C^2T)}{16C}.$$

Доведення. Моделі наведені в таблиці 2.1 для ясності.

Нехай $v^-=\left(Q_1^-,Q_2^-\right)$ — перша модель з розподілами Бернуллі Q_i^- для двох дій та $v^+=\left(Q_1^+,Q_2^+\right)$ — друга модель. Позначимо через $\mathbb{P}_{v_t^-}$ та $\mathbb{P}_{v_t^+}$ ймовірнісні міри для першої та другої моделі відповідно до означень 2.3 та 2.4 для усіх $t\in\{1,\ldots,T\}$. Так як максимум більше ніж середнє значення, маємо

$$\begin{aligned}\max\big(\mathbb{E}\left[L_{v^-}(T)\right],\mathbb{E}\left[L_{v^+}(T)\right]\big)&\geqslant\\&\geqslant\frac{1}{2}\big(\mathbb{E}\left[L_{v^-}(T)\right]+\mathbb{E}\left[L_{v^+}(T)\right]\big)=\\&=\frac{C}{4}\sum_{t=1}^{T}\big(\mathbb{P}_{v_t^-}\left(I_t=2\right)+\mathbb{P}_{v_t^+}\left(I_t=1\right)\big).\end{aligned}$$

Застосувавши нерівність Bretagnolle-Huber (лема 2.2) до останнього виразу, отримаємо наступну оцінку:

$$\max\left(\mathbb{E}\left[L_{v^-}(T)\right], \mathbb{E}\left[L_{v^+}(T)\right]\right) \geqslant \frac{C}{8}\sum_{t=1}^{T}\exp\left(-D_{\mathrm{KL}}\left(\mathbb{P}_{v_t^-}, \mathbb{P}_{v_t^+}\right)\right). \tag{2.4}$$

Використовуючи визначення для розподілу Бернуллі та декомпозицію спільних розподілів у розходженні Кульбака-Ляйблера до відособлених розподілів, отримаємо

$$D_{\mathrm{KL}}\left(\mathbb{P}_{v_t^-}, \mathbb{P}_{v_t^+}\right) = \sum_{i=1}^{2}\mathbb{E}_{v_t^-}\left[\sum_{s=1}^{t}\mathbb{1}_{\{I_s = i\}}\right] D_{\mathrm{KL}}\left(Q_i^-, Q_i^+\right),$$

де

$$D_{\mathrm{KL}}\left(Q_1^-, Q_1^+\right) = 0,$$

звідки маємо

$$\begin{aligned} &D_{\mathrm{KL}}\left(\mathbb{P}_{v_t^-}, \mathbb{P}_{v_t^+}\right) = \\ &\quad = \mathbb{E}_{v_t^-}\left[\sum_{s=1}^{t}\mathbb{1}_{\{I_s = 2\}}\right] D_{\mathrm{KL}}\left(Q_2^-, Q_2^+\right) = \\ &\quad = \mathbb{E}_{v_t^-}\left[\sum_{s=1}^{t}\mathbb{1}_{\{I_s = 2\}}\right]\left(\frac{1+C}{2}\log\left(\frac{1+C}{1-C}\right) + \frac{1-C}{2}\log\left(\frac{1-C}{1+C}\right)\right). \end{aligned}$$

Таким чином приходимо до наступної нерівності:

$$\begin{aligned} D_{\mathrm{KL}}\left(\mathbb{P}_{v_t^-}, \mathbb{P}_{v_t^+}\right) &= C\log\left(\frac{1+C}{1-C}\right)\mathbb{E}_{v_t^-}\left[\sum_{s=1}^{t}\mathbb{1}_{\{I_s = 2\}}\right] = \\ &= C\log\left(1 + \frac{2C}{1-C}\right)\mathbb{E}_{v_t^-}\left[\sum_{s=1}^{t}\mathbb{1}_{\{I_s = 2\}}\right] \leqslant \\ &\leqslant \frac{2C^2}{1-C}\,\mathbb{E}_{v_t^-}\left[\sum_{s=1}^{t}\mathbb{1}_{\{I_s = 2\}}\right]. \end{aligned}$$

Використовуючи обмеження $C \leqslant 1/2$, отримаємо з останньої нерівності наступну оцінку:

$$D_{\mathrm{KL}}\left(\mathbb{P}_{v_t^-}, \mathbb{P}_{v_t^+}\right) \leqslant 4C^2\,\mathbb{E}_{v_t^-}\left[\sum_{s=1}^{t}\mathbb{1}_{\{I_s = 2\}}\right]. \tag{2.5}$$

Отже, якщо підставити оцінку (2.5) в (2.4), маємо

$$\begin{aligned}
&\max\left(\mathbb{E}\left[L_{v^-}(T)\right], \mathbb{E}\left[L_{v^+}(T)\right]\right) \geqslant \\
&\quad \geqslant \frac{C}{8}\sum_{t=1}^{T}\exp\left(-4C^2\,\mathbb{E}_{v_t^-}\left[\sum_{s=1}^{t}\mathbb{1}_{\{I_s=2\}}\right]\right) \geqslant \\
&\quad \geqslant T\frac{C}{8}\exp\left(-4C^2\,\mathbb{E}_{v_T^-}\left[\sum_{t=1}^{T}\mathbb{1}_{\{I_t=2\}}\right]\right). \qquad (2.6)
\end{aligned}$$

Зазначимо тепер, що

$$\max\left(\mathbb{E}\left[L_{v^-}(T)\right], \mathbb{E}\left[L_{v^+}(T)\right]\right) \geqslant \mathbb{E}\left[L_{v^-}(T)\right] \geqslant \frac{C}{2}\,\mathbb{E}_{v_T^-}\left[\sum_{t=1}^{T}\mathbb{1}_{\{I_t=2\}}\right], \qquad (2.7)$$

оскільки втрати $C/2$ виникають при використанні першої моделі, коли обирається друга дія.

Отже, знову, враховуючи те, що максимум більший за середнє значення, підставимо (2.6) і (2.7) та отримаємо

$$\begin{aligned}
&\max\left(\mathbb{E}\left[L_{v^-}(T)\right], \mathbb{E}\left[L_{v^+}(T)\right]\right) \geqslant \\
&\quad \geqslant \frac{1}{2}\left(\frac{CT}{8}\exp\left(-4C^2\,\mathbb{E}_{v_T^-}\left[\sum_{t=1}^{T}\mathbb{1}_{\{I_t=2\}}\right]\right) + \frac{C}{2}\,\mathbb{E}_{v_T^-}\left[\sum_{t=1}^{T}\mathbb{1}_{\{I_t=2\}}\right]\right) = \\
&\quad = \frac{C}{4}\left(\frac{T}{4}\exp\left(-4C^2\,\mathbb{E}_{v_T^-}\left[\sum_{t=1}^{T}\mathbb{1}_{\{I_t=2\}}\right]\right) + \mathbb{E}_{v_T^-}\left[\sum_{t=1}^{T}\mathbb{1}_{\{I_t=2\}}\right]\right) \geqslant \\
&\quad \geqslant \min_{x=0,\ldots,T}\frac{C}{4}\left(\frac{T}{4}\exp\left(-4C^2x\right) + x\right) \geqslant \frac{\log(C^2T)}{16C}.
\end{aligned}$$

Що і треба було довести. □

Зауваження 2.5. Теореми 2.1 достатньо для знаходження оцінки очікуваних сукупних втрат у найгіршому випадку за означенням 2.11 у середовищі з двома діями зі спостереженнями, які мають розподіл Бернуллі. У такому випадку застосовується ідея перевірки статистичних гіпотез з використанням принципу компромісу і методу мінімаксу. Намір — показати складність на прикладі двох моделей, які одночасно є схожі та конкуренті. Опис надається у роботах [14, 68].

Таблиця 2.1

Параметри моделей з розподілом Бернуллі для пошуку нижньої границі у найгіршому випадку

Модель	Дія 1	Дія 2
v^+	$1/2$	$(1+C)/2$
v^-	$1/2$	$(1-C)/2$

2.3. Імовірнісні нерівності для асимптотичного аналізу верхньої границі втрат

У наступних розділах розглядаються асимптотично оптимальні стратегії у середовищі, яке представлене моделлю стохастичного багаторукого бандита зі спостереженнями, які мають бета-розподіл. Для пошуку верхньої границі ми будемо використовувати оцінки хвостів субгауссових випадкових величин та їх властивості ([16, 1]). Розглянемо поширені нерівності, щоб отримати властивості та оцінки для нашого випадку, які будемо використовувати в асимптотичному аналізі.

Означення 2.12 ([16])**.** Центрована випадкова величина η є σ-субгауссовою, якщо для всіх $\lambda \in \mathbb{R}$ існує $\sigma > 0$, що має місце нерівність

$$\mathbb{E}\left[\exp(\lambda\eta)\right] \leqslant \exp\left(\frac{\lambda^2\sigma^2}{2}\right).$$

Наслідок 2.1. *Нехай $\eta_1, \eta_2, \ldots, \eta_n$ — незалежні однаково розподілені σ-субгауссові випадкові величини. Тоді має місце наступна адитивна властивість:*

$$\eta_1 + \eta_2 + \cdots + \eta_n \text{ є } \sqrt{n\sigma^2}\text{-субгауссова випадкова величина.}$$

Доведення. Згідно з означенням 2.12 та незалежністю однаково розпо-

ділених величин маємо

$$\mathbb{E}\left[\exp\left(\lambda\left(\eta_1+\eta_2+\cdots+\eta_n\right)\right)\right]=\prod_{j=1}^{n}\mathbb{E}\left[\exp\left(\lambda\eta_j\right)\right]. \tag{2.8}$$

Оскільки експоненціальна функція монотонно зростає, отримаємо з останньої рівності наступну оцінку:

$$\mathbb{E}\left[\exp\left(\lambda\left(\eta_1+\eta_2+\cdots+\eta_n\right)\right)\right]\leqslant\prod_{j=1}^{n}\exp\left(\frac{\lambda^2\sigma^2}{2}\right)=\exp\left(\frac{\lambda^2\left(\sqrt{n\sigma^2}\right)^2}{2}\right),$$

що доводить наслідок. □

Наслідок 2.2. *Нехай η — центрована випадкова величина, яка має бета-розподіл. Тоді η є $1/2$-субгауссовою випадковою величиною.*

Доведення. Лема Хефдинга [48] стверджує, що для випадкової величини η з розподілом, який має носій функції $x\in[a,b]$, для всіх $\lambda\in\mathbb{R}$ має місце наступна нерівність:

$$\mathbb{E}\left[\exp(\lambda\eta)\right]\leqslant\exp\left(\lambda\,\mathbb{E}\left[\eta\right]+\frac{\lambda^2(b-a)^2}{8}\right).$$

Використовуючи центровану випадкову величину, яка має бета-розподіл, згідно з лемою Хефдинга отримаємо

$$\mathbb{E}\left[\exp(\lambda\eta)\right]\leqslant\exp\left(\frac{\lambda^2(1/2)^2}{2}\right),$$

що доводить наслідок. □

Підсумуємо властивості незалежних однаково розподілених σ-субгауссових випадкових величин η_j, які будуть корисними далі у даному та наступних розділах на додаток до наданих теорем:

- $\mathbb{E}\left[\eta_j\right]=0$;
- $\eta_1+\cdots+\eta_i+\cdots+\eta_n$ є $\sqrt{n\sigma^2}$-субгауссова випадкова величина;
- $C\eta_j$ — $|C|\sigma$-субгауссова випадкова величина, де $C\in\mathbb{R}$ — деяка стала.

Теорема 2.2 ([16]). *Нехай η — σ-субгауссова випадкова величина. Тоді для всіх $\varepsilon \geqslant 0$ виконується наступна нерівність:*

$$\mathbb{P}(\eta \geqslant \varepsilon) \leqslant \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right).$$

Наслідок 2.3. *Розглянемо незалежні однаково розподілені випадкові величини $\eta_1, \eta_2, \ldots, \eta_n$, які мають бета-розподіл з математичним сподіванням μ. Тоді для всіх $\varepsilon \geqslant 0$ мають місце наступні нерівності:*

$$\mathbb{P}\left(\frac{1}{n}\sum_{j=1}^{n}\eta_j \geqslant \mu + \varepsilon\right) \leqslant \exp\left(-2n\varepsilon^2\right)$$

та

$$\mathbb{P}\left(\frac{1}{n}\sum_{j=1}^{n}\eta_j \leqslant \mu - \varepsilon\right) \leqslant \exp\left(-2n\varepsilon^2\right).$$

Доведення. Згідно з властивостями σ-субгауссових випадкових величин та того, що $\eta_j - \mu$ є $1/2$-субгауссовою випадковою величиною, отримаємо наступну оцінку ймовірності з першої нерівності за допомогою теореми 2.2:

$$\mathbb{P}\left(\frac{1}{n}\sum_{j=1}^{n}\eta_j \geqslant \mu + \varepsilon\right) = \mathbb{P}\left(\frac{1}{n}\sum_{j=1}^{n}(\eta_j - \mu) \geqslant \varepsilon\right) \leqslant \exp\left(-\frac{\varepsilon^2}{2\left(\frac{\sqrt{n/4}}{n}\right)^2}\right),$$

що доводить наслідок для першої нерівності. Друга нерівність доводиться аналогічно. □

Зауваження 2.6. Замість оцінки хвостів субгауссових випадкових величин можна використовувати нерівність Чебишева:

$$\mathbb{P}\left(|\eta - \mathbb{E}[\eta]| \geqslant \varepsilon\right) \leqslant \frac{\mathrm{Var}[\eta]}{\varepsilon^2},$$

яка визначена для всіх $\varepsilon > 0$, де η — деяка випадкова величина з математичним сподіванням $\mathbb{E}[\eta]$ та дисперсією $\mathrm{Var}[\eta]$, які існують і є скінченними.

Приклад 2.2. Для нашого випадку з незалежними однаково розподіленими випадковими величинами $\eta_1, \eta_2, \ldots, \eta_n$ з середнім значенням μ та стандартним відхиленням σ оцінка Чебишева для всіх $\varepsilon > 0$ виглядає наступним чином:

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{j=1}^{n} \eta_j - \mu \right| \geqslant \varepsilon \right) \leqslant \frac{\sigma^2}{n\varepsilon^2}.$$

Оцінка $\frac{1}{n} \sum_{j=1}^{n} \eta_j$ є незміщеною оцінкою теоретичного середнього значення μ з дисперсією

$$\operatorname{Var}\left[\frac{1}{n} \sum_{j=1}^{n} \eta_j \right] = \mathbb{E}\left[\left(\frac{1}{n} \sum_{j=1}^{n} \eta_j - \mu \right)^2 \right] = \frac{\sigma^2}{n}.$$

Для бета-розподілу за деякими параметрами α і β маємо наступні характеристики:

$$\mu = \frac{\alpha}{\alpha + \beta} \in (0, 1),$$

та

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{(1-\mu)\mu}{\alpha+\beta+1} < (1-\mu)\mu \in (0, 0.25).$$

Тоді для пошуку верхньої границі у середовищі зі спостереженнями, які мають бета-розподіл, за допомогою нерівності Чебишева маємо наступну нерівність для всіх $\varepsilon > 0$:

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{j=1}^{n} \eta_j - \mu \right| \geqslant \varepsilon \right) \leqslant \frac{1}{4n\varepsilon^2}. \tag{2.9}$$

Наслідок 2.4. *Нехай $\eta_1, \eta_2, \ldots, \eta_n$ — незалежні однаково розподілені випадкові величини з деяким бета-розподілом з математичним сподіванням μ. Тоді для всіх $\varepsilon > 0$ оцінка хвостів за допомогою нерівності з наслідку 2.3*

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{j=1}^{n} \eta_j - \mu \right| \geqslant \varepsilon \right) \leqslant \exp\left(-2n\varepsilon^2\right)$$

є більш строгою, ніж за допомогою нерівності Чебишева (2.9)

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^{n}\eta_j - \mu\right| \geqslant \varepsilon\right) \leqslant \frac{1}{4n\varepsilon^2}.$$

Доведення. Щоб показати, що оцінка за допомогою нерівності з наслідку 2.3 є більш точною, ніж за допомогою нерівності Чебишева (2.9), достатньо показати, що для всіх $x > 0$ виконується

$$\exp(-x) \leqslant \frac{1}{2x} \leqslant \frac{1}{ex},$$

де $x = 2n\varepsilon^2$. Спочатку позбавимося експоненти та отримаємо

$$x - \log(x) - 1 \geqslant 0.$$

Зазначимо, що для всіх $x > 1$ виконується

$$(x - \log(x))' = 1 - 1/x > 0,$$

тобто маємо строго зростаючу функцію. Для всіх $x < 1$ маємо

$$(x - \log(x))' = 1 - 1/x < 0,$$

тобто строго спадну функцію, та $x = 1$ є єдиним розв'язком

$$x - \log(x) - 1 = 0.$$

Таким чином, оцінка за допомогою нерівності з наслідку 2.3 є більш строгою, ніж за допомогою нерівності Чебишева (2.9). □

Висновки до розділу 2

В цьому розділі був розглянутий послідовний розподіл ресурсів у стохастичному середовищі, яке представлене моделлю стохастичного багаторукого бандита. Надано математичну модель середовища відповідно до

розподілу винагород та їх характеристик. Надані означення класів середовищ зі спостереженнями, які мають Бернуллі та бета-розподіл відповідно. Ці означення будуть використовуватися далі в аналізі стратегій. Основними результатами даного розділу є:

- Отримано функцію очікуваних сукупних втрат з залежністю від неоптимальності дій.
- Наведені оцінки ефективності стратегії на основі рівномірного розподілу у стохастичному середовищі.
- Наведені приклади неоптимальних стратегій.
- Отримана нижня границя у найгіршому випадку для середовища з двома діями.
- Наведені ймовірнісні нерівності для асимптотичного аналізу втрат та отримані їх додаткові властивості.
- Отримані оцінки хвостів за допомогою субгауссових випадкових величин для середовища зі спостереженнями, які мають бета-розподіл. Доведено, що ці оцінки є більш строгими в порівнянні з нерівністю Чебишева.

РОЗДІЛ 3
АНАЛІЗ СТРАТЕГІЇ НА БАЗІ НАДІЙНОГО ІНТЕРВАЛУ

Даний розділ присвячений асимптотичному аналізу стратегії на базі надійного інтервалу у середовищі зі спостереженнями, які мають бета-розподіл. Наведено асимптотичний аналіз верхньої границі. Покращено оцінку ефективності стратегії для даного випадку. Проведено чисельні експерименти та наведені отримані дані. Результати даного розділу опубліковані у статті [34].

3.1. Попередні відомості та опис стратегії

Як згадувалось раніше, розглядається послідовний розподіл ресурсів у середовищі, яке представлене моделлю стохастичного багаторукого бандита. Моделюється послідовність прийняття рішення в умовах невизначеності у взаємодії між агентом та зовнішнім середовищем за допомогою обраної стратегії. Ця взаємодія відбувається протягом T кроків. На кожному кроці $t = 1, 2, \ldots, T$ агент обирає дію I_t із заданої множини $\{1, 2, \ldots, N\}$, у відповідь середовище видає винагороду $\xi_t \in \mathbb{R}_{\geqslant 0}$.

Стратегія на базі надійного інтервалу з використанням принципу «оптимізму в умовах невизначеності» була представлена у 2002 році у роботі [8], де автори показали, що вона є асимптотично оптимальною за означенням 2.10. Використовуючи розходження Кульбака-Ляйблера автори отримали наступний результат.

Теорема 3.1 ([8])**.** *Розглядається стохастичне середовище зі скінченим горизонтом T і кількістю дій N. Нехай носій функцій розподілів, пов'язаних з діями, є обмеженим. Припустимо, що перша дія є оптимальною без втрати*

загальності, тоді при використанні стратегії на базі надійного інтервалу має місце наступна нерівність:

$$\mathbb{E}\left[L(T)\right] \leqslant 8\left(\sum_{i=2}^{N}\frac{\log\left(T\right)}{\mu_1-\mu_i}\right)+\left(1+\frac{\pi^2}{3}\right)\sum_{i=2}^{N}\left(\mu_1-\mu_i\right).$$

Ця стратегія на кожному кроці $t \in \{1,\ldots,T\}$ для всіх дій $i \in \{1,\ldots,N\}$ обчислює значення індексу $U_i(t)$ на базі верхньої границі надійного інтервалу, яке з великою ймовірністю є завищеною оцінкою невідомого математичного сподівання розподілу, пов'язаного з дією i. На початку алгоритму стратегія вибирає кожну дію один раз, а потім на кожному наступному кроці t — дію з найбільшим значенням $U_i(t)$:

$$I_t = \begin{cases} t & \text{якщо } t \leqslant N, \\ \arg\max_{i=1,\ldots,N} U_i(t) & \text{інакше.} \end{cases}$$

Для отримання значення $U_i(t)$ нерівність оцінки хвостів представлена у вигляді надійного інтервалу та використана його верхня границя як сума вибіркового середнього та межі похибки. На прикладі нерівності Хефдинга [48] значення індексу виглядає наступним чином:

$$U_i(t) = \underbrace{\frac{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s=i\}}\xi_s}{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s=i\}}}}_{\text{використання}} + \underbrace{\sqrt{\lambda\frac{\log(t-1)}{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s=i\}}}}}_{\text{дослідження}}, \tag{3.1}$$

де λ — деяка стала, яка відповідає за темп дослідження. Це рівняння добре показує компроміс у виборі між дослідженням простору варіантів і використанням найоптимальнішого варіанту з раніше відомих для прийняття рішень у реальному часі в умовах невизначеності. Перша частина рівняння (3.1) (вибіркове середнє) відповідає за вибір кращої дії на даний час, а межа похибки — за дослідження. Вибір дії i відбувається, якщо її вибіркове середнє достатньо велике, що може свідчити про можливу оптимальність

дії, та/або, якщо межа похибки завелика — це може означати, що дія i недостатньо досліджена на горизонті T. Параметр λ допомагає контролювати цей баланс. Можна зазначити, що в (3.1) при

$$t \to \infty,$$
$$\sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = i\}} \ll \log(t-1),$$

маємо

$$\sqrt{\lambda \frac{\log(t-1)}{\sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = i\}}}} \to \infty,$$

що вказує на те, що неоптимальність дії i має як мінімум логарифмічну складність (нерівність (2.3)).

3.2. Алгоритм стратегії для середовища зі спостереженнями, які мають бета-розподіл

Розглянемо хвостову оцінку з наслідку 2.3 для незалежних однаково розподілених випадкових величин $\eta_1, \eta_2, \ldots, \eta_n$, які мають бета-розподіл з математичним сподіванням μ, у вигляді надійного інтервалу, де для всіх $\delta \in [0,1]$ з ймовірністю щонайменше $1-\delta$ маємо

$$\mu \in \left[\frac{1}{n}\sum_{j=1}^{n} \eta_j - \sqrt{\frac{\log(1/\delta)}{2n}}, \frac{1}{n}\sum_{j=1}^{n} \eta_j + \sqrt{\frac{\log(1/\delta)}{2n}} \right],$$

Звідси оцінка ймовірності для верхньої границі хвоста є

$$\mathbb{P}\left(\mu \geqslant \frac{1}{n}\sum_{j=1}^{n} \eta_j + \sqrt{\frac{\log(1/\delta)}{2n}} \right) \leqslant \delta \tag{3.2}$$

для всіх $\delta \in (0,1)$. Отже, треба обрати δ та показати, що, коли n є випадковою величиною у нерівності (3.2), δ все ще залишається оцінкою ймовірності, оскільки маємо випадкову величину $\sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = i\}}$ в (3.1).

Лема 3.1. *Розглянемо незалежні однаково розподілені випадкові величини* $\eta_1, \eta_2, \dots$ *на випадковому просторі* $(\Omega, \mathcal{F}, \mathbb{P})$*, які мають бета-розподіл з математичним сподіванням* μ*. Нехай* $n : \Omega \to \{1, 2, \dots\}$ *— деяка випадкова величина. Тоді для всіх* $\delta \in (0, 1)$ *маємо*

$$\mathbb{P}\left(\mu \geqslant \frac{1}{n}\sum_{j=1}^{n}\eta_j + \sqrt{\frac{\log(1/\delta)}{2n}}\right) \leqslant \delta.$$

Доведення. Використовуючи правило повного математичного сподівання та нерівність (3.2), де n є сталою, отримаємо

$$\begin{aligned}
&\mathbb{P}\left(\mu \geqslant \frac{1}{n}\sum_{j=1}^{n}\eta_j + \sqrt{\frac{\log(1/\delta)}{2n}}\right) = \\
&= \sum_{i=1}^{\infty}\mathbb{E}\left[\mathbb{1}_{\{n=i\}}\mathbb{1}_{\left\{\sum_{j=1}^{i}(\mu-\eta_j)\geqslant\sqrt{\frac{i\log(1/\delta)}{2}}\right\}}\right] = \\
&= \sum_{i=1}^{\infty}\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{n=i\}}\mathbb{1}_{\left\{\sum_{j=1}^{i}(\mu-\eta_j)\geqslant\sqrt{\frac{i\log(1/\delta)}{2}}\right\}}\right] \,\middle|\, n\right] = \\
&= \sum_{i=1}^{\infty}\mathbb{E}\left[\mathbb{1}_{\{n=i\}}\,\mathbb{E}\left[\mathbb{1}_{\left\{\sum_{j=1}^{i}(\mu-\eta_j)\geqslant\sqrt{\frac{i\log(1/\delta)}{2}}\right\}}\right] \,\middle|\, n\right] \leqslant \\
&\leqslant \sum_{i=1}^{\infty}\mathbb{E}\left[\mathbb{1}_{\{n=i\}}\delta\right] = \delta,
\end{aligned}$$

що і треба було довести. □

На скінченному горизонті T для середовища зі спостереженнями (ξ_t), які мають бета-розподіл, з оцінки (3.2) та леми 3.1 маємо наступний індекс для дії i, використовуючи стратегію на базі надійного інтервалу:

$$U_i(t) = \frac{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s=i\}}\xi_s}{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s=i\}}} + \sqrt{\frac{\log(T)}{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s=i\}}}}, \tag{3.3}$$

де вірогідний рівень обрано $\delta = 1/T^2$.

Отже, пропонується наступний алгоритм.

Алгоритм 3.1. *Алгоритм стратегії на базі надійного інтервалу для середовища зі спостереженнями, які мають бета-розподіл.* Розглядається стоха-

стичне середовище зі скінченим горизонтом T і кількістю дій N. Кожна дія $i \in \{1, \dots, N\}$ має бета-розподіл з невідомим математичним сподіванням μ_i. Вибираючи дію I_t, модель виконує відбір ξ_t з розподілу, пов'язаного з дією I_t та, як результат, реалізація вибірки стає доступною для стратегії.

Крок 1. Покласти $t = 1$.

Крок 2. Якщо $t \leqslant N$, то покласти $I_t = t$ та перейти до кроку 5.

Крок 3. Для кожної дії $i \in \{1, \dots, N\}$ покласти

$$U_i(t) = \frac{\sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = i\}} \xi_s}{\sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = i\}}} + \sqrt{\frac{\log(T)}{\sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = i\}}}}.$$

Крок 4. Призначити $I_t = \arg\max_{i=1,\dots,N} U_i(t)$.

Крок 5. Виконати відбір ξ_t з розподілу, пов'язаного з дією I_t.

Крок 6. Якщо $t > T$, то закінчити виконання алгоритму. Інакше — збільшити t на 1 та перейти до кроку 2.

3.3. Асимптотичний аналіз верхньої границі втрат

Використовуючи загальні методи асимптотичного аналізу подібних алгоритмів на базі надійного інтервалу, описаних в роботах [14, 56], отримаємо наступну верхню границю для нашого випадку за допомогою оцінки хвостів субгауссових випадкових величин.

Теорема 3.2. *Розглядається стохастичне середовище зі скінченим горизонтом T і кількістю дій N. Кожна дія $i \in \{1, \dots, N\}$ має бета-розподіл з невідомим математичним сподіванням μ_i. Припустимо, що перша дія є оптимальною без втрати загальності. Тоді при використанні стратегії на базі надійного інтервалу за алгоритмом 3.1 має місце наступна нерівність:*

$$\mathbb{E}\left[L(T)\right] \leqslant 2 \sum_{i=2}^{N} (\mu_1 - \mu_i) + \frac{1}{2} \sum_{i=2}^{N} \frac{\log(T)}{\mu_1 - \mu_i}.$$

Доведення. Згідно з алгоритмом 3.1 для оцінки математичного сподівання з послідовності винагород $(\xi_t)_{t=1}^T$ виразимо верхню границю дії i на кроці t наступним чином:

$$U_i(t) = \begin{cases} \infty & \text{якщо } \sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = i\}} = 0, \\ \frac{\sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = i\}} \xi_s}{\sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = i\}}} + \sqrt{\frac{\log(T)}{\sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = i\}}}} & \text{інакше.} \end{cases}$$

Для того, щоб отримати верхню границю очікуваних сукупних втрат, будемо використовувати визначення втрат з залежністю від неоптимальності дій з леми 2.1

$$\mathbb{E}\left[L(T)\right] = \sum_{i=1}^{N} \max_{j=1,\dots,N} (\mu_j - \mu_i)\, \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{I_t = i\}}\right]. \tag{3.4}$$

Наша мета полягає в знаходженні оцінки очікуваної кількості виборів неоптимальної дії $i > 1$ за весь горизонт T.

Так як, мета стратегії — максимізація очікуваної сукупної винагороди $\sum_{t=1}^{T} \xi_t$, що веде до необхідності визначення оптимальної дії якомога раніше, отже, до мінімізації обсягу вибірки $\sum_{t=1}^{T} \mathbb{1}_{\{I_t = i\}}$ для неоптимальної дії $i > 1$.

Нехай B_i — подія, за якої виконуються наступні дві умови:

1. Значення оцінки верхньої границі $U_i(T)$ дії $i > 1$ з обсягом вибірки c_i є меншим ніж математичне сподівання μ_1 оптимальної дії.
2. Оцінка верхньої границі $U_1(t)$ оптимальної дії на кожному кроці не є недооціненою відносно її математичного сподівання μ_1.

Отже, маємо

$$B_i = \left\{ U_i(T) < \mu_1 \;\middle|\; \sum_{t=1}^{T} \mathbb{1}_{\{I_t = i\}} = c_i \right\} \cap \left\{ \mu_1 < \min_{t=1,\dots,T} U_1(t) \right\}.$$

Ми зацікавленні мати високоймовірну подію B_i з найменшим можливим розміром вибірки c_i. Використовуючи правило повного математичного

сподівання для очікуваного обсягу вибірки

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}\right]= \\
&=\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}\,\middle|\,B_i\right]\mathbb{P}(B_i)+\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}\,\middle|\,B_i^c\right]\mathbb{P}(B_i^c)\leqslant \\
&\leqslant (c_i+T\,\mathbb{P}(B_i^c)),
\end{aligned} \tag{3.5}$$

отримаємо наступну оцінку втрат, підставляючи (3.5) в (3.4):

$$\begin{aligned}
\mathbb{E}[L(T)]&=\sum_{i=1}^{N}\max_{j=1,\dots,N}(\mu_j-\mu_i)\,\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}\right]= \\
&\leqslant\sum_{i=2}^{N}(\mu_1-\mu_i)\,(c_i+\mathbb{P}(B_i^c)\,T),
\end{aligned} \tag{3.6}$$

де $B_i^c=\Omega\setminus B_i$ та за визначенням

$$B_i^c=\left\{U_i(T)\geqslant\mu_1\,\middle|\,\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}=c_i\right\}\cup\left\{\mu_1\geqslant\min_{t=1,\dots,T}U_1(t)\right\}. \tag{3.7}$$

Знайдемо оцінку ймовірності другої події в (3.7) за допомогою оцінок хвостів субгауссових випадкових величин з леми 2.3:

$$\begin{aligned}
&\mathbb{P}\left(\mu_1\geqslant\min_{t=1,\dots,T}U_1(t)\right)\leqslant \\
&\leqslant\sum_{t=1}^{T}\mathbb{P}(\mu_1\geqslant U_1(t))= \\
&=\sum_{t=1}^{T}\mathbb{P}\left(\mu_1\geqslant\frac{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s=1\}}\xi_s}{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s=1\}}}+\sqrt{\frac{\log(T)}{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s=1\}}}}\right)\leqslant \\
&\leqslant\sum_{t=1}^{T}\exp\left(-2\left(\sqrt{\frac{\log(T)}{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s=1\}}}}\right)^2\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s=1\}}\right)= \\
&=\sum_{t=1}^{T}\exp\left(-\log(T^2)\right)=\frac{1}{T}.
\end{aligned}$$

Підставивши цю оцінку в (3.6), отримаємо

$$\begin{aligned}\mathbb{E}\left[L(T)\right] &\leqslant \sum_{i=2}^{N}\left(\mu_1-\mu_i\right)\left(c_i+\mathbb{P}\left(B_i^c\right)T\right)\leqslant\\ &\leqslant \sum_{i=2}^{N}\left(\mu_1-\mu_i\right)\left(c_i+1+\mathbb{P}\left(U_i(T)\geqslant\mu_1\,\Big|\,\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}=c_i\right)T\right).\end{aligned} \tag{3.8}$$

Отже, щоб позбутися останньої випадковості

$$\mathbb{P}\left(U_i(T)\geqslant\mu_1\,\Big|\,\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}=c_i\right)$$

для оцінки очікуваних сукупних втрат з залежністю від неоптимальності дій, скористаємося означенням неоптимальності дії з (2.2) та лемою 2.3:

$$\begin{aligned}&\mathbb{P}\left(U_i(T)\geqslant\mu_1\,\Big|\,\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}=c_i\right)=\\ &=\mathbb{P}\left(U_i(T)\geqslant\mu_i+\left(\mu_1-\mu_i\right)\,\Big|\,\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}=c_i\right)=\\ &=\mathbb{P}\left(\frac{\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}\xi_t}{\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}}+\sqrt{\frac{\log(T)}{\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}}}\geqslant\mu_i+\left(\mu_1-\mu_i\right)\,\Big|\,\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}=c_i\right)\\ &=\mathbb{P}\left(\frac{\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}\xi_t}{\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}}\geqslant\mu_i+\left(\mu_1-\mu_i-\sqrt{\frac{\log(T)}{\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}}}\right)\,\Big|\,\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}=c_i\right)\\ &\leqslant\exp\left(-2c_i\left(\mu_1-\mu_i-\sqrt{\frac{\log(T)}{c_i}}\right)^2\right).\end{aligned} \tag{3.9}$$

Щоб обрати c_i, мінімізуємо оцінку сукупних втрат через розв'язок наступного рівняння:

$$\exp\left(-2c_i\left(\mu_1-\mu_i-\sqrt{\frac{\log(T)}{c_i}}\right)^2\right)=\frac{1}{T},$$

де ми отримуємо

$$c_i = \frac{\left(3 \pm 2\sqrt{2}\right)\log(T)}{2\left(\mu_1 - \mu_i\right)^2}.$$

Візьмемо найменший за значенням розв'язок та зробимо наближення до цілого:

$$c_i = \frac{\log(T)}{2\left(\mu_1 - \mu_i\right)^2} \tag{3.10}$$

Тепер залишається підставити оцінку ймовірності (3.9) зі значенням $1/T$ та (3.10) в (3.8):

$$\begin{aligned}
\mathbb{E}\left[L(T)\right] &\leqslant \sum_{i=2}^{N}\left(\mu_1 - \mu_i\right)\left(c_i + 1 + \mathbb{P}\left(U_i(T) \geqslant \mu_1 \,\middle|\, \sum_{t=1}^{T}\mathbb{1}_{\{I_t = i\}} = c_i\right)T\right) = \\
&= \sum_{i=2}^{N}(\mu_1 - \mu_i)\left(\frac{\log(T)}{2\left(\mu_1 - \mu_i\right)^2} + 1 + 1\right) = \\
&= 2\sum_{i=2}^{N}\left(\mu_1 - \mu_i\right) + \frac{1}{2}\sum_{i=2}^{N}\frac{\log(T)}{\mu_1 - \mu_i}.
\end{aligned}$$

Що і треба було довести. □

Таким чином, ми отримали верхню границю очікуваних сукупних втрат для стратегії на базі надійного інтервалу для середовища зі спостереженнями, які мають бета-розподіл. Ця оцінка є асимптотично оптимальною за означенням 2.10.

3.4. Чисельні експерименти

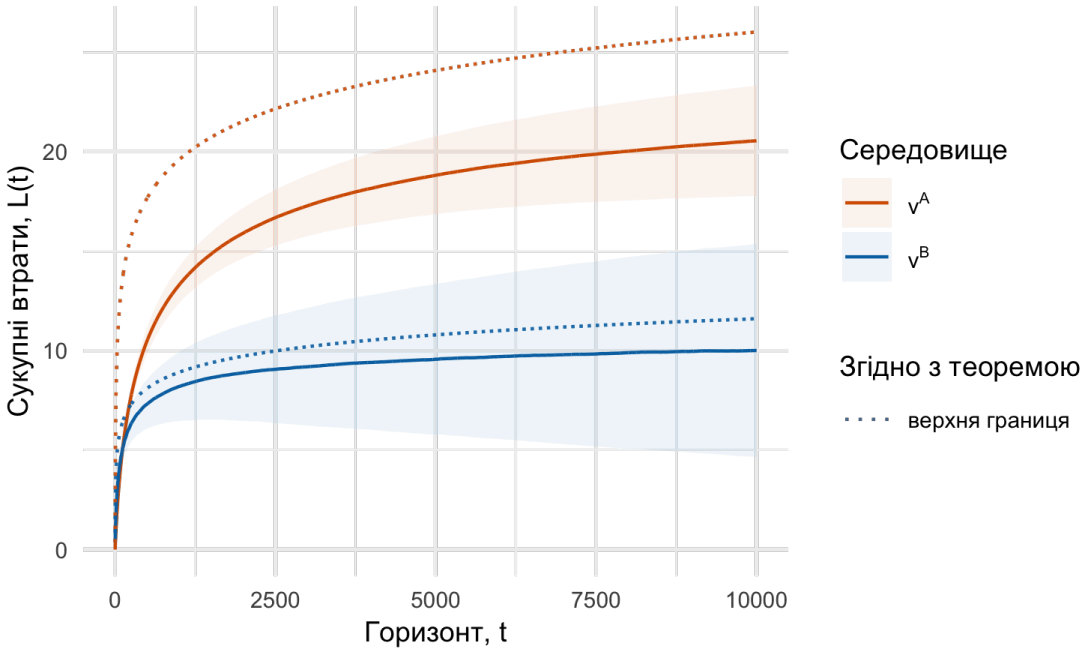
У цьому підрозділі представлені результати емпіричних тестів для стратегії на базі надійного інтервалу у стохастичному середовищі зі спостереженнями, які мають бета-розподіл. Для цього було розроблено програмне забезпечення [28, 29] з імплементацією алгоритму 3.1 для середовища класу $\mathcal{V}^{\text{Beta}}$ з означення 2.2. Більш детальний опис математичного моделювання надається у розділі 6.

Мета експериментів — показати властивості та асимптотичну поведінку наведеної стратегії у середовищах з різними параметрами. Результати всіх експериментів агреговані з 10000 незалежних тестів і зображені на рисунках 3.1 та 3.2. На цих графіках можна побачити зворотну залежність між часом потрібним на пошук оптимальної дії та різницею математичних сподівань між діями й пряму залежність між часом пошуку оптимальності та кількістю дій в середовищі. Також на рисунках наведено верхні границі згідно з теоремою 3.2, які підтверджують отримані теоретичні результати. Теоретична верхня границя досить близько апроксимує емпіричні результати. Використані параметри середовищ наведені в таблиці 3.1.

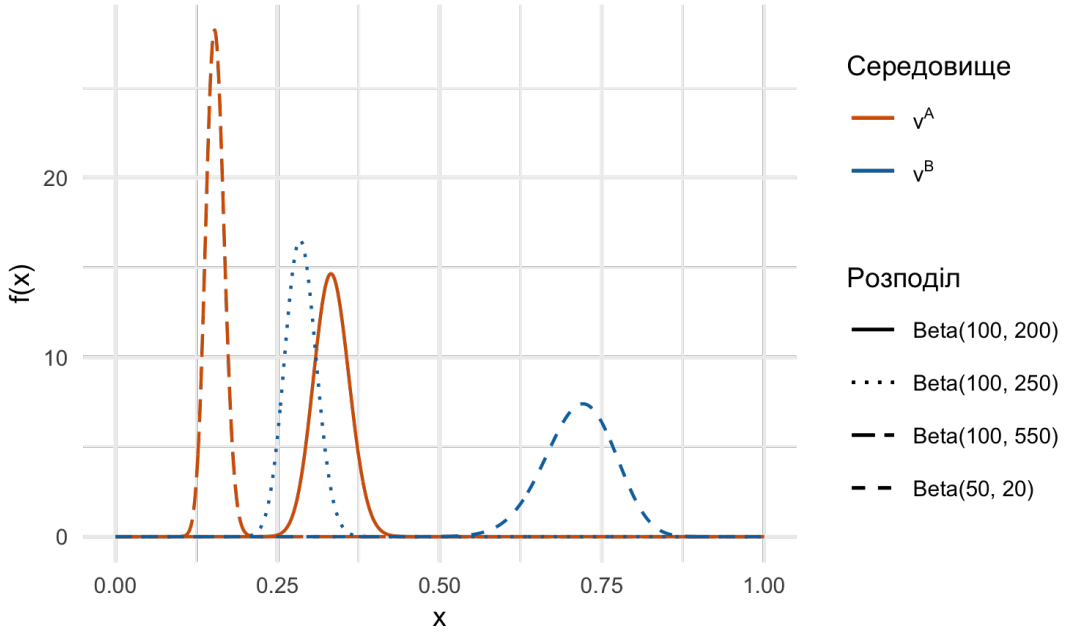
Таблиця 3.1

Параметри дій середовищ з бета-розподілом для експериментів зі стратегією на базі надійного інтервалу

Модель $\boldsymbol{v} \in \mathcal{V}^{\text{Beta}}$	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6
v^A	0.33	0.15				
v^B	0.29	0.71				
v^C	0.33	0.71				
v^D	0.29	0.15	0.56	0.71		
v^E	0.29	0.15	0.56	0.31	0.39	0.71

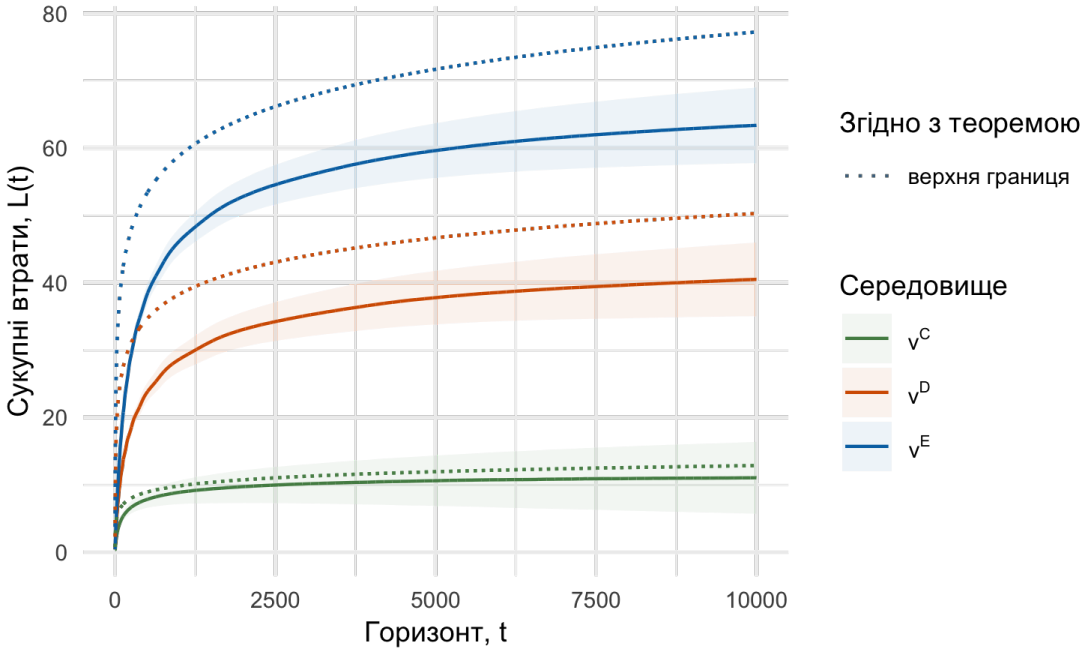


(а) Сукупні втрати на кожному кроці

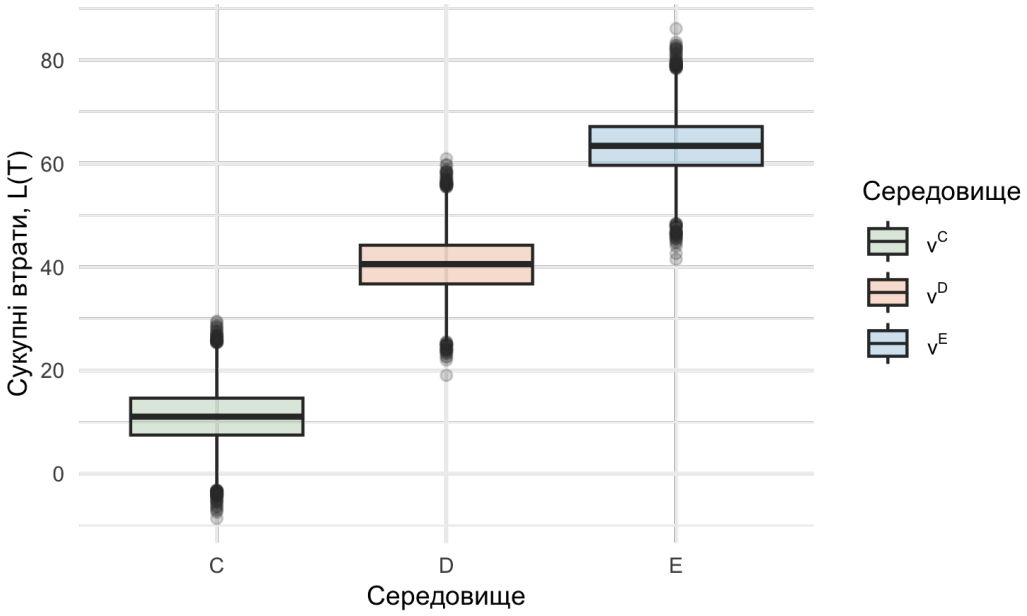


(б) Розподіли дій

Рис. 3.1: Результати експериментів у середовищі зі спостереженнями, які мають бета-розподіл, для стратегії на базі надійного інтервалу. Демонструють різницю в асимптотичній поведінці в залежності від параметрів розподілів дій. Середовища A та B мають по дві дії. Різниця математичних сподівань першого середовища дорівнює 0.18, другого — 0.43.



(а) Сукупні втрати на кожному кроці



(б) Сукупні втрати на горизонті T

Рис. 3.2: Результати експериментів у стохастичному середовищі зі спостереженнями, які мають бета-розподіл, для стратегії на базі надійного інтервалу. Демонструють різницю в асимптотичній поведінці в залежності від кількості дій: (C) дві, (D) чотири та (E) шість. Усі варіанти середовищ (C, D та E) мають однакове математичне сподівання оптимальної дії.

Висновки до розділу 3

В цьому розділі була досліджена стратегія на базі надійного інтервалу для випадку середовища зі спостереженнями, які мають бета-розподіл. Зокрема, були отримані наступні результати:

- Адаптовано алгоритм для середовища зі спостереженнями, які мають бета-розподіл.
- Отримана асимптотична оцінка очікуваних сукупних втрат.
- Проведено чисельні експерименти, які підтверджують теоретичні результати.

Серед напрямків подальшого дослідження виділимо наступні:

- Дослідити вплив вибору значення вірогідного рівня, який було покладено $\delta = 1/T^2$ в алгоритмі 3.1. Даний вибір був продиктований загальними рекомендаціями для подібних алгоритмів, отже, залишається питання оптимальності для нашого випадку.
- Розглянути альтернативний алгоритм розроблений в роботах [41, 53], де замість нерівності Чебишева використовується нерівність Чернова [19] та індекс виглядає наступним чином:

$$U_i(t) = \sup_{q \in [0,1]} \left\{ q : D_{\mathrm{KL}} \left(\frac{\sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = i\}} \xi_s}{\sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = i\}}}, q \right) \leqslant \frac{\log(t)}{\sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = i\}}} \right\}. \tag{3.11}$$

  Цей варіант стратегії є також асимптотично оптимальним за означенням 2.10. Для реалізації даного алгоритму потрібно розв'язувати задачу оптимізації в (3.11). Оскільки розходження Кульбака-Ляйблера є опуклою функцією, ми можемо скористатися методом бісекції чи Ньютона.

РОЗДІЛ 4
АНАЛІЗ БАЄСОВОЇ СТРАТЕГІЇ

У даному розділі розглядається баєсова стратегія у середовищі зі спостереженнями, які мають бета-розподіл. Розроблено алгоритм для нашого випадку на базі існуючого алгоритму для середовища зі спостереженнями, які мають розподіл Бернуллі. Отримано оцінку ефективності стратегії з точки зору баєсового аналізу. Описано підхід з марковськими процесами прийняття рішень. Показані результати чисельних експериментів. Опис баєсової стратегії та марковських процесів прийняття рішень, які розглядаються у даному розділі, опубліковано у статтях [27, 36].

4.1. Попередні відомості та опис стратегії

Розглянемо невідомий параметр $\boldsymbol{\theta}$ як випадкову величину з відбору вибірки з деякого апріорного розподілу Π. На кожному кроці $t = 1, 2, \ldots, T$ стратегія обирає дію I_t із заданої множини відповідно до пов'язаних з ними розподілів

$$\boldsymbol{\upsilon_\theta} = (\upsilon_{\theta_1}, \upsilon_{\theta_2}, \ldots, \upsilon_{\theta_N}) .$$

У відповідь середовище видає винагороду $\xi_t \in \mathbb{R}_{\geqslant 0}$. Позначимо апостеріорний розподіл параметра $\boldsymbol{\theta}$ на кроці t як

$$\Pi_t (\boldsymbol{\theta}) = \mathcal{L} (\boldsymbol{\theta} \mid I_1, \xi_1, \ldots, I_t, \xi_t) ,$$

де апостеріорний розподіл параметра θ_i після m спостережень дії i є

$$\pi_t^i = \mathcal{L} (\theta_i | \xi_{i,1}, \ldots, \xi_{i,m}) ,$$

де $\xi_{i,s}$ — це s-спостереження при виборі дії i. У баєсових стратегіях вибір дії заснований на поточному апостеріорному розподілі параметра $\boldsymbol{\theta}$. В ро-

боті [65] була запропонована асимптотично оптимальна баєсова стратегія з наступним алгоритмом.

Алгоритм 4.1. *Алгоритм баєсової стратегії.* Розглядається стохастичне середовище зі скінченим горизонтом T і кількістю дій N. Кожна дія $i \in \{1, \ldots, N\}$ має розподіл з існуючим та скінченним математичним сподіванням μ_i. Вибираючи дію I_t, модель виконує відбір ξ_t з розподілу, пов'язаного з дією I_t та, як результат, реалізація вибірки стає доступною для стратегії.

Крок 1. Покласти $t = 1$.

Крок 2. Якщо $t \leqslant N$, то покласти $I_t = t$ та перейти до кроку 5.

Крок 3. Зробити оцінку параметра $\hat{\boldsymbol{\theta}} = \mathcal{L}\left(\cdot \mid I_1, \xi_1, \ldots, I_{t-1}, \xi_{t-1}\right)$.

Крок 4. Призначити $I_t = \arg\max_{i=1,\ldots,N} \hat{\theta}_i$.

Крок 5. Виконати відбір ξ_t з розподілу, пов'язаного з дією I_t.

Крок 6. Якщо $t > T$, то закінчити виконання алгоритму. Інакше — збільшити t на 1 та перейти до кроку 2.

Таким чином, на кожному кроці маємо баєсове прийняття рішення та відбір з кожного апріорного розподілу з обранням дії з найбільшим значенням для поточної оцінки невідомого параметра.

Використовуючи класичний аналіз була отримана наступна оцінка одночасно у двох роботах [49, 3].

Теорема 4.1 ([49, 3])**.** *Розглядається стохастичне середовище зі спостереженнями, які мають розподіл Бернуллі, зі скінченим горизонтом T та кількістю дій N. Припустимо, що перша дія є оптимальною без втрати загальності. Візьмемо рівномірний розподіл для апріорного розподілу Π_0. Тоді при використанні баєсової стратегії за алгоритмом 4.1 має місце наступна нерівність:*

$$\mathbb{E}_{\boldsymbol{\theta}}\left[L(T)\right] \leqslant \left(\sum_{i=2}^{N} \frac{1}{(\mu_1 - \mu_i)^2}\right)^2 \log(T).$$

Для загального випадку була отримана наступна нерівність ([15]):

$$\mathbb{E}_{\boldsymbol{\theta}}\left[L(T)\right] \leqslant 14\sqrt{NT}. \tag{4.1}$$

У наступному підрозділі ми використаємо отримані результати для розробки алгоритму для середовища зі спостереженнями, які мають бета-розподіл.

З точки зору баєсового аналізу, були отримані наступні важливі для нашого випадку результати.

Теорема 4.2 ([14])**.** *Розглядається стохастичне середовище зі скінченим горизонтом T і кількістю дій N. Нехай розподіли дій мають носій функцій $x \in [0,1]$. Тоді при використанні баєсової стратегії за алгоритмом 4.1 існує апріорний розподіл Π_0 для якого має місце наступна нерівність:*

$$\mathbb{E}_{\mathbf{\Pi_0}}\left[L(T)\right] \geqslant \frac{1}{20}\sqrt{NT}.$$

4.2. Алгоритм стратегії для середовища зі спостереженнями, які мають розподіл Бернуллі

У цьому підрозділі ми побудуємо алгоритм для середовища зі спостереженнями, які мають бета-розподіл, на базі алгоритму 4.1 з використанням розподілу Бернуллі. Покажемо, що результати з теореми 4.1 релевантні для нашого випадку.

Розглянемо наступну ієрархічну модель для середовища зі спостереженнями (ξ_t), які мають бета-розподіл:

$$\begin{aligned} \eta_t|\xi_t &\sim \text{Bern}(\xi_t), \quad t = 1, \ldots, T, \\ \xi_t &\sim \text{Beta}(\alpha_{I_t}, \beta_{I_t}). \end{aligned} \tag{4.2}$$

Отже, маємо внутрішнє середовище зі спостереженнями (η_t), які мають розподіл Бернуллі. Для кожної дії $i \in \{1, \ldots, N\}$ маємо послідовність

$$\{\, \eta_t \in \{0,1\} : t \in \{1, \ldots, T\} \wedge I_t = i \,\}$$

випадкових величин Бернуллі з ймовірністю успіху θ_i. Звичайним спряженим апріором у такому випадку є щільність бета-розподілу. Припустимо, що в нас m реалізацій вибірки (x_m) для дії i, з яких маємо s успішних результатів (подія $\{1\}$). Тоді функція правдоподібності для θ_i є

$$p(x_1, \ldots, x_m|\theta_i) = \theta_i^s(1-\theta_i)^{m-s}.$$

Використовуючи щільність бета-розподілу

$$p(\theta_i) = \frac{1}{\mathrm{B}(\alpha_0, \beta_0)}\theta_i^{\alpha_0-1}\left(1-\theta_i\right)^{\beta_0-1},$$

для всіх $\theta_i \in [0,1]$ і бета-функцію

$$\mathrm{B}(\alpha_0, \beta_0) = \int_0^1 x^{\alpha_0-1}(1-x)^{\beta_0-1}\,dx,$$

отримаємо наступну апостеріорну ймовірність:

$$p(\theta_i|x_1, \ldots, x_m) = \frac{p(x_1, \ldots, x_m|\theta_i)p(\theta_i)}{C},$$

де C — це стала для виконання умови нормування та

$$C = \mathrm{B}(\alpha_0 + s, \beta_0 + m - s).$$

Задамо $\alpha_0 = \beta_0 = 1$ і таким чином отримаємо рівномірний розподіл $\mathrm{Unif}(0,1)$ на початку алгоритму:

$$\pi_0^i = \mathrm{Beta}(1,1).$$

Отже, на кроці t виводиться наступний апостеріорний розподіл:

$$\pi_t^i = \mathrm{Beta}\left(\sum_{s=1}^{t}\mathbb{1}_{\{I_s=i\}}\mathbb{1}_{\{\eta_s=1\}}+1, \sum_{s=1}^{t}\mathbb{1}_{\{I_s=i\}} - \sum_{s=1}^{t}\mathbb{1}_{\{I_s=i\}}\mathbb{1}_{\{\eta_s=1\}}+1\right).$$

Таким чином, маємо наступний алгоритм.

Алгоритм 4.2. *Алгоритм баєсової стратегії для середовища зі спостереженнями, які мають бета-розподіл.* Розглядається стохастичне середовище

зі скінченим горизонтом T і кількістю дій N. Кожна дія $i \in \{1, \dots, N\}$ має бета-розподіл з невідомим математичним сподіванням μ_i. Вибираючи дію I_t, модель виконує відбір ξ_t з розподілу, пов'язаного з дією I_t та, як результат, реалізація вибірки стає доступною для стратегії.

Крок 1. Покласти $t = 1$.

Крок 2. Для кожної дії $i \in \{1, \dots, N\}$ покласти $\alpha_i = 1, \beta_i = 1$.

Крок 3. Для кожної дії $i \in \{1, \dots, N\}$ виконати відбір

$$\hat{\theta}_i \sim \text{Beta}(\alpha_i, \beta_i).$$

Крок 4. Призначити $I_t = \arg\max_{i=1,\dots,N} \hat{\theta}_i$.

Крок 5. Виконати відбір ξ_t з розподілу, пов'язаного з дією I_t.

Крок 6. Виконати відбір $\eta_t \sim \text{Bern}(\xi_t)$.

Крок 7. Покласти

$$\alpha_{I_t} = \alpha_{I_t} + \eta_t,$$
$$\beta_{I_t} = \beta_{I_t} + (1 - \eta_t).$$

Крок 8. Якщо $t > T$, то закінчити виконання алгоритму. Інакше — збільшити t на 1 та перейти до кроку 3.

Наслідок 4.1. *Розглядається стохастичне середовище зі скінченим горизонтом T, кількістю дій N та спостереженнями, які мають бета-розподіл. Припустимо, що перша дія є оптимальною без втрати загальності. Візьмемо рівномірний розподіл для апріорного розподілу Π_0. Тоді при використанні баєсової стратегії за алгоритмом 4.2 має місце наступна нерівність:*

$$\mathbb{E}_{\boldsymbol{\theta}}\left[L(T)\right] \leqslant \left(\sum_{i=2}^{N} \frac{1}{(\mu_1 - \mu_i)^2}\right)^2 \log(T).$$

Доведення. Маємо ієрархічну модель для середовища зі спостереженнями (ξ_t), які мають бета-розподіл:

$$\eta_t | \xi_t \sim \text{Bern}(\xi_t), \quad t = 1, \dots, T,$$

$$\xi_t \sim \text{Beta}(\alpha_{I_t}, \beta_{I_t}).$$

Зазначимо, що ймовірність успішного результату ($\eta_t = 1$) дорівнює математичному сподіванню μ_{I_t} бета-розподілу поточної дії I_t:

$$\begin{aligned}\mathbb{P}\left(\eta_t = 1\right) &= \mathbb{P}\left(\eta_t = 1, 0 \leqslant \xi_t \leqslant 1\right) = \\ &= \int\limits_0^1 \xi_t \frac{1}{\text{B}(\alpha_{I_t}, \beta_{I_t})} \xi_t^{\alpha_{I_t}-1} \left(1-\xi_t\right)^{\beta_{I_t}-1} d\xi_t = \mu_{I_t}.\end{aligned}$$

Таким чином маємо властивості середовища зі спостереженнями, які мають розподіл Бернуллі з математичним сподіванням $(\mu_i : i \in \{1, \ldots, N\})$ як і в роботі [49], що доводить наслідок. □

Зауважимо, що ми можемо виводити апостеріорний розподіл на кроці t без використання розподілу Бернуллі наступним чином:

$$\pi_t^i = \text{Beta}\left(\sum_{s=1}^{t} \mathbb{1}_{\{I_s = i\}} \xi_s + 1, \sum_{s=1}^{t} \mathbb{1}_{\{I_s = i\}} - \sum_{s=1}^{t} \mathbb{1}_{\{I_s = i\}} \xi_s + 1\right),$$

але це потребує окремого аналізу. Наведемо цей алгоритм і зробимо порівняння з алгоритмом 4.2 за допомогою математичного моделювання.

Алгоритм 4.3. *Алгоритм баєсової стратегії без вибірки з розподілу Бернуллі.* Розглядається стохастичне середовище зі скінченим горизонтом T і кількістю дій N. Кожна дія $i \in \{1, \ldots, N\}$ має бета-розподіл з невідомим математичним сподіванням μ_i. Вибираючи дію I_t, модель виконує відбір ξ_t з розподілу, пов'язаного з дією I_t та, як результат, реалізація вибірки стає доступною для стратегії.

Крок 1. Покласти $t = 1$.

Крок 2. Для кожної дії $i \in \{1, \ldots, N\}$ покласти $\alpha_i = 1, \beta_i = 1$.

Крок 3. Для кожної дії $i \in \{1, \ldots, N\}$ виконати відбір

$$\hat{\theta}_i \sim \text{Beta}(\alpha_i, \beta_i).$$

Крок 4. Призначити $I_t = \arg\max_{i=1,\ldots,N} \hat{\theta}_i$.

Крок 5. Виконати відбір ξ_t з розподілу, пов'язаного з дією I_t.

Крок 6. Покласти

$$\alpha_{I_t} = \alpha_{I_t} + \xi_t,$$
$$\beta_{I_t} = \beta_{I_t} + (1 - \xi_t).$$

Крок 7. Якщо $t > T$, то закінчити виконання алгоритму. Інакше — збільшити t на 1 та перейти до кроку 3.

4.3. Баєсів аналіз верхньої границі втрат

Використовуючи загальні методи асимптотичного аналізу подібних баєсових алгоритмів, описаних в роботах [68, 10], отримаємо наступну верхню границю для нашого випадку.

Теорема 4.3. *Розглядається стохастичне середовище зі скінченим горизонтом T, кількістю дій N та спостереженнями, які мають бета-розподіл. Припустимо, що перша дія є оптимальною без втрати загальності. Тоді при використанні баєсової стратегії за алгоритмом 4.1 має місце наступна нерівність для баєсових очікуваних сукупних втрат:*

$$\mathbb{E}_{\mathbf{\Pi_0}}\left[L(T)\right] \leqslant 8N + 4\sqrt{NT\log(T)}.$$

Доведення. Використаємо результати з оцінки хвостів субгауссових випадкових величин для випадкових величин з бета-розподілом з попереднього розділу, де ми отримали індекс верхньої границі надійного інтервалу (3.3) з вірогідним рівнем $\delta = 1/T^2$,

$$U_i(t) = \frac{\sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = i\}} \xi_s}{\sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = i\}}} + \sqrt{\frac{\log(T)}{\sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = i\}}}},$$

для середовища зі спостереженнями (ξ_t).

Нагадаємо, що маємо фільтрацію (2.1)

$$\mathcal{F}_t = \sigma\left(I_1, \xi_1, I_2, \xi_2, \ldots, I_t, \xi_t\right),$$

де $\mathcal{F}_s \subset \mathcal{F}_t, \forall s < t$, $s, t \in \{1, \ldots, T\}$. Тоді баєсові очікувані сукупні втрати з означення 2.7 можна виразити наступним чином:

$$\begin{aligned}
\mathbb{E}_{\mathbf{\Pi_0}}\left[L^{\kappa}(T)\right] &= \mathbb{E}\left[\sum_{t=1}^{T}(\mu_1 - \mu_{I_t})\right] = \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}\left[\mu_1 - \mu_{I_t} \,|\, \mathcal{F}_{t-1}\right]\right] = \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}\left[\mu_1 - U_{I_t}(t) + U_{I_t}(t) - \mu_{I_t} \,|\, \mathcal{F}_{t-1}\right]\right].
\end{aligned} \tag{4.3}$$

З означення 2.4 маємо наступне твердження для алгоритму 4.1:

$$\mathbb{P}\left(\arg\max_{j=1,\ldots,N}\mu_j = \cdot \,|\, \mathcal{F}_{t-1}\right) = \mathbb{P}\left(I_t = \cdot \,|\, \mathcal{F}_{t-1}\right),$$

а тому з (4.3)

$$\begin{aligned}
\mathbb{E}_{\mathbf{\Pi_0}}\left[L^{\kappa}(T)\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}\left[\mu_1 - U_1(t) + U_{I_t}(t) - \mu_{I_t} \,|\, \mathcal{F}_{t-1}\right]\right] = \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\left(\mathbb{E}\left[\mu_1 - U_1(t) \,|\, \mathcal{F}_{t-1}\right] + \mathbb{E}\left[U_{I_t}(t) - \mu_{I_t} \,|\, \mathcal{F}_{t-1}\right]\right)\right] = \\
&= \mathbb{E}\left[\sum_{t=1}^{T}(\mu_1 - U_1(t)) + \sum_{t=1}^{T}(U_{I_t}(t) - \mu_{I_t})\right].
\end{aligned} \tag{4.4}$$

Нехай B — випадкова подія, за якої виконується умова

$$\left|\frac{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s = i\}}\xi_s}{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s = i\}}} - \mu_i\right| < \sqrt{\frac{\log(T)}{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s = i\}}}}$$

для всіх $i \in \{1, \ldots, N\}$ та $t \in \{1, \ldots, T\}$. Разом з (4.4) та використовуючи

правило повного математичного сподівання маємо

$$\begin{aligned}\mathbb{E}_{\mathbf{\Pi_0}}\left[L^{\kappa}(T)\right] = \mathbb{E}\left[\sum_{t=1}^{T}(\mu_1 - U_1(t)) + \sum_{t=1}^{T}(U_{I_t}(t) - \mu_{I_t}) \,\Big|\, B\right]\mathbb{P}\left(B\right) + \\ + \mathbb{E}\left[\sum_{t=1}^{T}(\mu_1 - U_1(t)) + \sum_{t=1}^{T}(U_{I_t}(t) - \mu_{I_t}) \,\Big|\, B^c\right]\mathbb{P}\left(B^c\right),\end{aligned} \tag{4.5}$$

де $B^c = \Omega \setminus B$. Зазначимо, що перша сума у першому доданку (4.5) є від'ємною за умови B, тоді оцінимо перший доданок наступним чином:

$$\begin{aligned}&\mathbb{P}\left(B\right)\mathbb{E}\left[\sum_{t=1}^{T}(U_{I_t}(t) - \mu_{I_t}) \,\Big|\, B\right] = \\ &= \mathbb{P}\left(B\right)\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{1}_{\{I_t = i\}}(U_i(t) - \mu_i) \,\Big|\, B\right] = \\ &= \mathbb{P}\left(B\right)\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{1}_{\{I_t = i\}}\left(\frac{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s = i\}}\xi_s}{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s = i\}}} + \sqrt{\frac{\log(T)}{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s = i\}}}} - \mu_i\right) \,\Big|\, B\right] \leqslant \\ &\leqslant \mathbb{P}\left(B\right)\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{1}_{\{I_t = i\}}\sqrt{\frac{4\log(T)}{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s = i\}}}} \,\Big|\, B\right] = \\ &= \mathbb{P}\left(B\right)\mathbb{E}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{1}_{\{I_t = i\}}\sqrt{\frac{4\log(T)}{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s = i\}}}} \,\Big|\, B\right] \leqslant \\ &\leqslant \mathbb{E}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{1}_{\{I_t = i\}}\sqrt{\frac{4\log(T)}{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s = i\}}}} \,\Big|\, B\right] \leqslant \\ &\leqslant \mathbb{E}\left[\sum_{i=1}^{N}\int_{0}^{\sum_{s=1}^{T}\mathbb{1}_{\{I_s = i\}}}\sqrt{\frac{4\log(T)}{x}}\,dx \,\Big|\, B\right] = \\ &= \mathbb{E}\left[\sum_{i=1}^{N}\sqrt{16\log(T)\sum_{s=1}^{T}\mathbb{1}_{\{I_s = i\}}} \,\Big|\, B\right] \leqslant \\ &= \sqrt{16NT\log(T)}.\end{aligned} \tag{4.6}$$

Для першої суми у другому доданку за умови B^c з (4.5) маємо

$$\mathbb{E}\left[\sum_{t=1}^{T}(\mu_1 - U_1(t)) \,\Big|\, B^c\right]\mathbb{P}\left(B^c\right) =$$

$$= \mathbb{E}\left[\sum_{t=1}^{T}\left(\mu_1 - \frac{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s=1\}}\xi_s}{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s=1\}}} - \sqrt{\frac{\log(T)}{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s=1\}}}}\right)\,\middle|\, B^c\right]\mathbb{P}\left(B^c\right) \leqslant$$

$$\leqslant 2\,\mathbb{E}\left[\sum_{t=1}^{T}\left(\mu_1 - \frac{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s=1\}}\xi_s}{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s=1\}}}\right)\,\middle|\, B^c\right]\mathbb{P}\left(B^c\right) \leqslant$$

$$\leqslant 2T\,\mathbb{P}\left(B^c\right). \tag{4.7}$$

Далі, використаємо оцінку ймовірності для верхньої границі хвоста (3.2)

$$\mathbb{P}\left(\mu \geqslant \frac{1}{n}\sum_{j=1}^{n}\eta_j + \sqrt{\frac{\log(1/\delta)}{2n}}\right) \leqslant \delta, \quad \forall\delta \in (0,1),$$

для випадкових величин (η_j), які мають бета-розподіл. Маємо T кроків і N дій, тоді за визначенням події B та вірогідним рівнем $\delta = 1/T^2$ отримаємо з нерівності (4.7) наступне:

$$\mathbb{E}\left[\sum_{t=1}^{T}(\mu_1 - U_1(t))\,\middle|\, B^c\right]\mathbb{P}\left(B^c\right) \leqslant 4N. \tag{4.8}$$

Аналогічно для другої суми у другому доданку за умови B^c з (4.5) маємо

$$\mathbb{E}\left[\sum_{t=1}^{T}(U_{I_t}(t) - \mu_{I_t})\,\middle|\, B^c\right]\mathbb{P}\left(B^c\right) =$$

$$= \mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s=I_t\}}\xi_s}{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s=I_t\}}} + \sqrt{\frac{\log(T)}{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s=I_t\}}}} - \mu_{I_t}\right)\,\middle|\, B^c\right]\mathbb{P}\left(B^c\right) \leqslant$$

$$\leqslant 2\,\mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s=I_t\}}\xi_s}{\sum_{s=1}^{t-1}\mathbb{1}_{\{I_s=I_t\}}} - \mu_{I_t}\right)\,\middle|\, B^c\right]\mathbb{P}\left(B^c\right) \leqslant$$

$$\leqslant 2T\,\mathbb{P}\left(B^c\right) \leqslant 4N. \tag{4.9}$$

А отже, з (4.6), (4.8) та (4.9) випливає, що оцінка втрат (4.5) набуває

$$\mathbb{E}_{\mathbf{\Pi_0}}\left[L^{\kappa}(T)\right] \leqslant \sqrt{16NT\log(T)} + 8N,$$

що і треба було довести. □

4.4. Марковські процеси прийняття рішень

В продовження баєсового аналізу, розглянемо підхід до описання моделі за допомогою марковських процесів прийняття рішень (МППР).

У загальному випадку модель багаторукого бандита може бути описана за допомогою МППР ([63]). Позначимо через $\Delta(B)$ множину всіх імовірнісних розподілів групи B, через S та $\mathcal{I}$ — множини станів та дій відповідно. Нехай P — деяке стохастичне ядро, яке задає ймовірність $P(s, i, s')$ того, що дія i в стані s на кроці t призведе до стану s' на кроці $t+1$. Для кожного стану $s \in S$ та дії $i \in \mathcal{I}$ маємо $P(s, i, \cdot) \in \Delta(S)$. Нехай R — деяке стохастичне ядро, яке задає стохастичні винагороди $R(s, i, s', \cdot) \in \Delta([0, 1])$. Тоді можна задати МППР четвіркою $(S, \mathcal{I}, P, R)$.

Коли агент знаходиться у стані $s \in S$ та вибирає дію $i \in \mathcal{I}$, відбувається перехід до нового стану $s' \sim P(\cdot|s, i)$ та отримання винагороди $r \sim R(\cdot|s, i)$. Мета агента — знайти стратегію (функцію, яка вказує, яку дію вибрати відповідно до поточного стану), яка максимізує очікувану винагороду у МППР з невідомими параметрами. Ця взаємодія агента з середовищем зображена на рисунку 4.1.

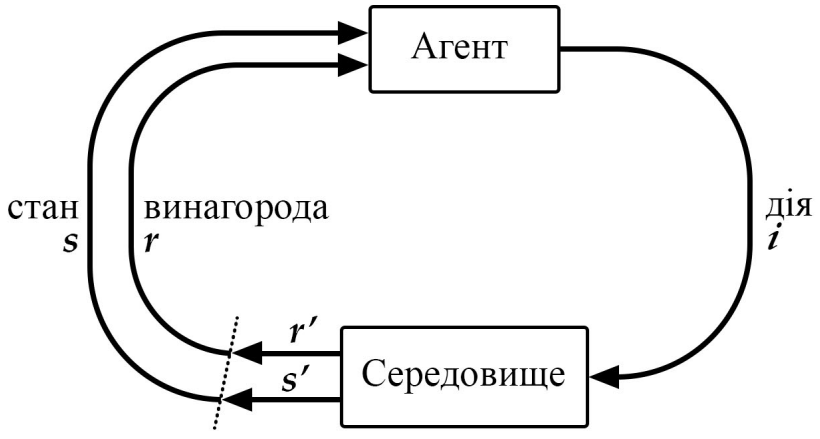


Рис. 4.1: Взаємодія агента зі стохастичним середовищем.

Стаціонарна стохастична модель багаторукого бандита може бути опи-

сана як МППР з одним станом s_0, тоді для стохастичного ядра, яке задає винагороди, маємо $R(\cdot|s_0, i) = v_{\theta_i}$.

За допомогою МППР побудуємо модель баєсового багаторукого бандита, розглянутого у попередньому підрозділі, де поточний стан — це поточний апостеріорний розподіл Π, а відбір $\boldsymbol{\theta}$ робиться з апріорного розподілу Π_0. Таким чином, агент обирає дію i у стані Π, отримує винагороду $r \sim v_{\theta_i}$, та, враховуючи нові спостереження, виводить новий апостеріорний розподіл Π'.

Пошук оптимальної стратегії у цій МППР моделі є можливим методами динамічного програмування, де функція цінностей V для нескінченого горизонту та деякого коефіцієнта знецінювання $\gamma \in (0, 1]$ має вигляд

$$V(\Pi) = \mathbb{E}_{\mathbf{\Pi_0}} \left[\sum_{t=1}^{\infty} \gamma^{t-1} \xi_t \right],$$

чи для скінченного горизонту T

$$V(\Pi, T) = \mathbb{E}_{\mathbf{\Pi_0}} \left[\sum_{t=1}^{T} \xi_t \right].$$

Нехай A_i^r — деякий оператор перетворення апостеріорного розподілу:

$$\Pi' = A_i^r(\Pi).$$

Тоді, з теорії марковських процесів прийняття рішень маємо оптимальну стратегію з наступним рекурсивним алгоритмом:

$$V(\Pi) = \max_{i \in \mathcal{I}} \Big(\mathbb{E}_{\theta_i \sim \pi_i} [\mu(\theta_i)] + \gamma \, \mathbb{E}_{r \sim v_{\theta_i}, \theta_i \sim \pi_i} [V(A_i^r(\Pi))] \Big),$$

та для скінченного горизонту:

$$V(\Pi, t) = \max_{i \in \mathcal{I}} \Big(\mathbb{E}_{\theta_i \sim \pi_i} [\mu(\theta_i)] + \gamma \, \mathbb{E}_{r \sim v_{\theta_i}, \theta_i \sim \pi_i} [V(A_i^r(\Pi), t-1)] \Big),$$

за умови, що $V(\Pi, 0) = 0$.

Знаходження оптимальної стратегії у МППР моделі може вимагати забагато обчислення у зв'язку з потенційно великим простором станів. Розв'язки для випадку дворукого бандита були наведені у роботах [39, 12].

4.5. Чисельні експерименти

У цьому підрозділі представлені результати емпіричних тестів для баєсової стратегії у стохастичному середовищі зі спостереженнями, які мають бета-розподіл. Для цього було розроблено програмне забезпечення [28, 29] з імплементацією алгоритмів 4.2 і 4.3 для середовища класу $\mathcal{V}^{\text{Beta}}$ з означення 2.2. Більш детальний опис математичного моделювання надається у розділі 6.

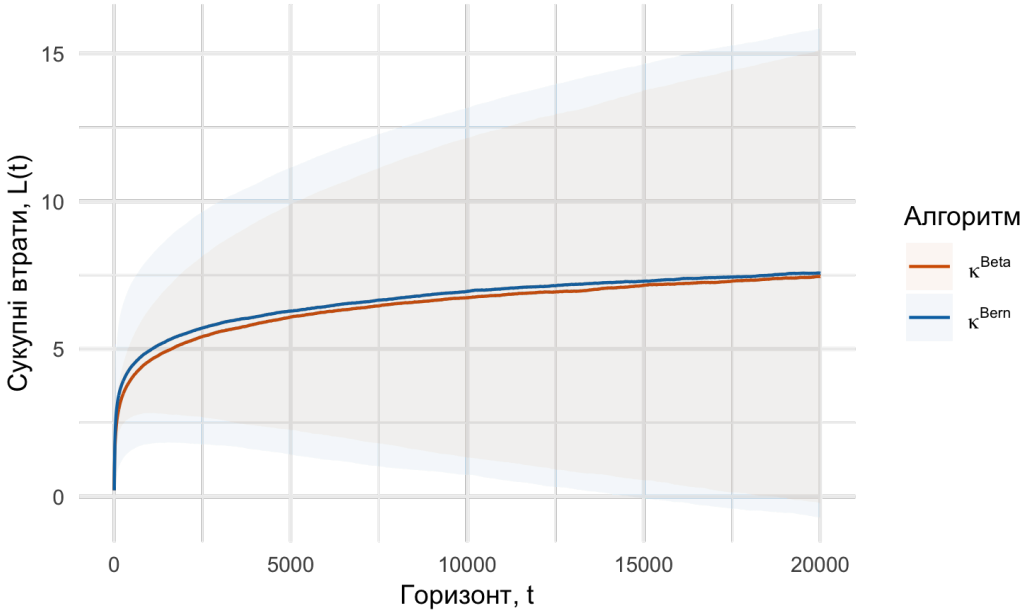
Мета експериментів — показати властивості та асимптотичну поведінку наведеної стратегії у середовищах з різними параметрами та порівняти алгоритм 4.2 з 4.3. Результати всіх експериментів агреговані з 10000 незалежних тестів і зображені на рисунку 4.2.

Як і у випадку стратегії на базі надійного інтервалу, на цих графіках можна побачити зворотну залежність між часом потрібним на пошук оптимальної дії та різницею математичних сподівань між діями й пряму залежність між часом пошуку оптимальності та кількістю дій в середовищі. Також зазначимо, що цей алгоритм має значно більше розсіювання з експериментів (відхилення) в порівнянні зі стратегією на базі надійного інтервалу. Використані параметри середовищ наведені в таблиці 4.1.

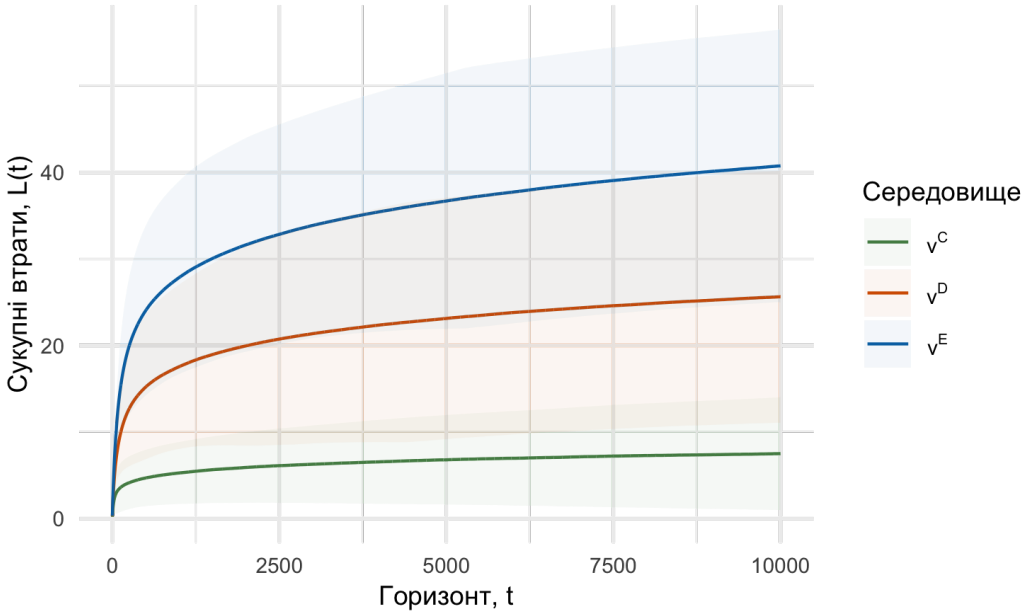
Таблиця 4.1

Параметри дій середовищ з бета-розподілом для експериментів з використанням баєсової стратегії

Модель $\boldsymbol{v} \in \mathcal{V}^{\text{Beta}}$	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6
v^B	0.29	0.71				
v^C	0.33	0.71				
v^D	0.29	0.15	0.56	0.71		
v^E	0.29	0.15	0.56	0.31	0.39	0.71



(а) Сукупні втрати алгоритмів 4.2 (κ^{Bern}) та 4.3 (κ^{Beta})



(б) Сукупні втрати алгоритму 4.2

Рис. 4.2: Результати експериментів у середовищі зі спостереженнями, які мають бета-розподіл для баєсової стратегії.

(а) Результати алгоритмів 4.2 (κ^{Bern}) та 4.3 (κ^{Beta}) у середовищі (B).

(б) Результати алогритма 4.2 в залежності від кількості дій: (C) дві, (D) чотири та (E) шість; з однаковим матем. сподіванням оптимальної дії.

Висновки до розділу 4

В цьому розділі була досліджена баєсова стратегія для випадку середовища зі спостереженнями, які мають бета-розподіл. Зокрема, були отримані наступні результати:

- Розроблено алгоритм для середовища зі спостереженнями, які мають бета-розподіл.
- Отримано асимптотичну оцінку очікуваних сукупних втрат з залежністю від неоптимальності дій.
- Отримано баєсову асимптотичну оцінку очікуваних сукупних втрат без залежності від неоптимальності дій та припущень щодо апріорного розподілу.
- Проведено чисельні експерименти, які демонструють пошук оптимальних дій за допомогою розглянутих алгоритмів.

Серед напрямків подальшого дослідження виділимо наступні:

- Дослідити асимптотичну поведінку, отриману з математичного моделювання, для чого проаналізувати відхилення, яке є значно більшим в порівнянні зі стратегією на базі надійного інтервалу.
- Зробити асимптотичний аналіз алгоритму 4.3, отримати оцінку очікуваних сукупних втрат з залежністю від неоптимальності дій.

РОЗДІЛ 5
АНАЛІЗ ЖАДІБНОЇ СТРАТЕГІЇ

Даний розділ присвячений асимптотичному аналізу жадібної стратегії у середовищі зі спостереженнями, які мають бета-розподіл. Наведено асимптотичний аналіз верхньої границі. Отримано оцінку ефективності стратегії для випадку з двома діями та оптимальним вибором кількості досліджень простору варіантів. Отримано оцінку ефективності стратегії у загальному випадку. Наведені результати проведених чисельних експериментів. Значна частина результатів даного розділу опублікована у статті [26].

5.1. Попередні відомості та опис стратегії

Перші приклади жадібної стратегії з'явилися у роботі [65], де було показано, що за деяких умов втрати можуть бути сублінійними. Ця стратегія характеризується фіксованою кількістю C досліджень кожної дії в першій фазі дослідження простору варіантів і вибором емпірично кращої дії для використання у другій фазі. Алгоритм вибору дії I_t на кроці t:

$$I_t = \begin{cases} (t \bmod N) + 1 & \text{якщо } t \leqslant CN, \\ \arg\max_{i=1,\dots,N} \frac{\sum_{s=1}^{t} \mathbb{1}_{\{I_s = i\}} \xi_s}{\sum_{s=1}^{t} \mathbb{1}_{\{I_s = i\}}} & \text{інакше,} \end{cases}$$

для середовища зі спостереженнями (ξ_t).

Алгоритм стратегії виглядає наступним чином.

Алгоритм 5.1. *Алгоритм жадібної стратегії з фіксованою кількістю C досліджень кожної дії.* Розглядається стохастичне середовище зі скінченим горизонтом T і кількістю дій N. Кожна дія $i \in \{1, \dots, N\}$ має деякий

розподіл з невідомим математичним сподіванням μ_i. Вибираючи дію I_t, модель виконує відбір ξ_t з розподілу, пов'язаного з дією I_t та, як результат, реалізація вибірки стає доступною для стратегії.

Крок 1. Покласти C та $t = 1$.

Крок 2. Якщо $t \leqslant CN$, то покласти

$$I_t = (t \bmod N) + 1$$

та перейти до кроку 4.

Крок 3. Призначити

$$I_t = \underset{i=1,\ldots,N}{\arg\max} \frac{\sum_{s=1}^{t} \mathbb{1}_{\{I_s = i\}} \xi_s}{\sum_{s=1}^{t} \mathbb{1}_{\{I_s = i\}}}.$$

Крок 4. Виконати відбір ξ_t з розподілу, пов'язаного з дією I_t.

Крок 5. Якщо $t > T$, то закінчити виконання алгоритму. Інакше — збільшити t на 1 та перейти до кроку 2.

Якщо горизонт T відомий наперед, ми можемо мінімізувати верхню границю для отримання сублінійних втрат. Для моделі з розподілом Бернуллі автори роботи [68] показали наступну оцінку:

$$\mathbb{E}\left[L(T)\right] \leqslant T^{2/3} \left(N \log T\right)^{1/3},$$

де C була обрана наступним чином:

$$C = \left(T/N\right)^{2/3} \left(\log T\right)^{1/3}.$$

Варіацією цієї стратегії без залежності від кількості кроків T є стратегія, яка займається дослідженням та використанням протягом усього горизонту. Задається додатковий параметр $0 < \varepsilon < 1$, та на кожному кроці з ймовірністю ε відбувається дослідження простору варіантів, вибираючи рівномірно випадково, та з ймовірністю $1 - \varepsilon$ використовується найкраща дія за вибірковим середнім:

$$I_t = \begin{cases} i \sim \text{Unif}(\{1, N\}) & \text{з ймовірністю } \varepsilon, \\ \arg\max_{i=1,\ldots,N} \frac{\sum_{s=1}^{t} \mathbb{1}_{\{I_s = i\}} \xi_s}{\sum_{s=1}^{t} \mathbb{1}_{\{I_s = i\}}} & \text{інакше,} \end{cases}$$

де $\mathrm{Unif}(\{a, b\})$ — дискретний рівномірний розподіл з параметрами $b \geqslant a$. Втрати цієї стратегії лінійні, так як ми змушенні продовжувати досліджувати варіанти. Границя знизу становить щонайменше

$$\mathbb{E}\left[L(T)\right] \geqslant \left(\varepsilon \frac{1}{N} \sum_{i=1}^{N} \max_{j=1,\dots,N} \left(\mu_j - \mu_i\right)\right) T.$$

В [8] було показано, що можливо отримати логарифмічну складність, якщо замість сталого значення ε використовувати спадну послідовність (ε_t). Для вибору кроку послідовності потрібно знати наперед значення

$$\min_{i=2,\dots,N} \left(\mu_1 - \mu_i\right),$$

де перша дія є оптимальною без втрати загальності.

5.2. Асимптотичний аналіз верхньої границі втрат у випадку двох дій

Для випадку середовища з двома діями позначимо неоптимальність як

$$\Delta\mu = |\mu_1 - \mu_2|.$$

Теорема 5.1. *Розглядається стохастичне середовище зі скінченим горизонтом T, двома діями та неоптимальністю $\Delta\mu$. Кожна дія i має бета-розподіл з невідомим математичним сподіванням μ_i. Тоді при використанні жадібної стратегії за алгоритмом 5.1 має місце наступна нерівність:*

$$\mathbb{E}\left[L(T)\right] \leqslant C\Delta\mu + (T - 2C)\Delta\mu \exp\left(-C\left(\Delta\mu\right)^2\right).$$

Доведення. Припустимо, що перша дія є оптимальною без втрати загальності. Для аналізу втрат з залежністю від неоптимальності дій (лема 2.1) нам потрібно оцінити очікувану кількість вибору другої дії алгоритмом за всі T кроків. Ймовірність того, що друга дія матиме найбільше

вибіркове середнє після $2C$ кроків досліджень є

$$\mathbb{P}\left(\frac{\sum_{t=1}^{2C}\mathbb{1}_{\{I_t=2\}}\xi t}{\sum_{t=1}^{2C}\mathbb{1}_{\{I_t=2\}}}\geqslant\frac{\sum_{t=1}^{2C}\mathbb{1}_{\{I_t=1\}}\xi t}{\sum_{t=1}^{2C}\mathbb{1}_{\{I_t=1\}}}\right).$$

Тоді, відповідно до протоколу алгоритму 5.1, кожна дія вибирається рівно C разів під час дослідження і $(T-2C)$ разів під час використання з ймовірністю отримання найбільшого середнього, тобто маємо

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}_{\{I_t=2\}}\right]= \\
&\quad= C+(T-2C)\,\mathbb{P}\left(\frac{\sum_{t=1}^{2C}\mathbb{1}_{\{I_t=2\}}\xi t}{\sum_{t=1}^{2C}\mathbb{1}_{\{I_t=2\}}}\geqslant\frac{\sum_{t=1}^{2C}\mathbb{1}_{\{I_t=1\}}\xi t}{\sum_{t=1}^{2C}\mathbb{1}_{\{I_t=1\}}}\right)= \\
&\quad= C+(T-2C)\,\mathbb{P}\left(\frac{\sum_{t=1}^{2C}\mathbb{1}_{\{I_t=2\}}\xi t}{\sum_{t=1}^{2C}\mathbb{1}_{\{I_t=2\}}}-\mu_2-\left(\frac{\sum_{t=1}^{2C}\mathbb{1}_{\{I_t=1\}}\xi t}{\sum_{t=1}^{2C}\mathbb{1}_{\{I_t=1\}}}-\mu_1\right)\geqslant\Delta\mu\right)= \\
&\quad= C+(T-2C)\,\mathbb{P}\left(\frac{\sum_{t=1}^{2C}\mathbb{1}_{\{I_t=2\}}(\xi_t-\mu_2)}{\sum_{t=1}^{2C}\mathbb{1}_{\{I_t=2\}}}-\frac{\sum_{t=1}^{2C}\mathbb{1}_{\{I_t=1\}}(\xi_t-\mu_1)}{\sum_{t=1}^{2C}\mathbb{1}_{\{I_t=1\}}}\geqslant\Delta\mu\right),
\end{aligned}\tag{5.1}$$

де $\xi_t-\mu_{I_t}$ є центрованою випадковою величиною з бета-розподілом. Згідно з лемою 2.3 і властивостями незалежних однаково розподілених субгаусових випадкових величин отримаємо наступну оцінку ймовірності з (5.1):

$$\begin{aligned}
&\mathbb{P}\left(\frac{\sum_{t=1}^{2C}\mathbb{1}_{\{I_t=2\}}(\xi_t-\mu_2)}{\sum_{t=1}^{2C}\mathbb{1}_{\{I_t=2\}}}-\frac{\sum_{t=1}^{2C}\mathbb{1}_{\{I_t=1\}}(\xi_t-\mu_1)}{\sum_{t=1}^{2C}\mathbb{1}_{\{I_t=1\}}}\geqslant\Delta\mu\right)\leqslant \\
&\qquad\leqslant\exp\left(-\frac{(\Delta\mu)^2}{2\left(1/\sqrt{2C}\right)^2}\right)=\exp\left(-C\,(\Delta\mu)^2\right).
\end{aligned}$$

Звідси разом з (5.1) та використовуючи лему 2.1 маємо

$$\begin{aligned}
\mathbb{E}\left[L(T)\right]&=\sum_{i=1}^{2}(\mu_1-\mu_i)\,\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}\right]\leqslant \\
&\leqslant\Delta\mu\left(C+(T-2C)\exp\left(-C\,(\Delta\mu)^2\right)\right),
\end{aligned}$$

що і треба було довести. □

Верхня границя з теореми 5.1 добре відображає складність знаходження балансу між дослідженням і використанням. Якщо C занадто мала, тоді стратегія недостатньо займається дослідженням простору варіантів та ймовірність вибору неоптимальної дії збільшується, як і значення правої частини нерівності. Якщо C завелика, ми збільшуємо ліву частину нерівності, яка безпосередньо відповідає за втрати від дослідження.

Наслідок 5.1. *При оптимізації кількості досліджень простору варіантів жадібна стратегія має наступну оцінку втрат у середовищі зі спостереженнями, які мають бета-розподіл, зі скінченим горизонтом T, двома діями та неоптимальністю $\Delta\mu$:*

$$\mathbb{E}\left[L(T)\right] \leqslant \frac{1}{\Delta\mu} + \frac{1}{\Delta\mu}\log\left(T\left(\Delta\mu\right)^2\right).$$

Доведення. Мінімізуємо оцінку сукупних втрат з теореми 5.1

$$\begin{aligned}\mathbb{E}\left[L(T)\right] &\leqslant C\Delta\mu + (T-2C)\Delta\mu\exp\left(-C\left(\Delta\mu\right)^2\right) \leqslant \\ &\leqslant C\Delta\mu + T\Delta\mu\exp\left(-C\left(\Delta\mu\right)^2\right)\end{aligned}$$

за допомогою

$$C = \frac{1}{\left(\Delta\mu\right)^2}\log\left(T\left(\Delta\mu\right)^2\right),$$

що доводить наслідок. □

Отже, з наслідку 5.1 маємо наступну формулу знаходження кількості досліджень простору варіантів для оптимізації ефективності стратегії для випадку з двома діями:

$$C = \frac{1}{\left(\Delta\mu\right)^2}\log\left(T\left(\Delta\mu\right)^2\right). \tag{5.2}$$

5.3. Асимптотичний аналіз верхньої границі втрат у загальному випадку

Теорема 5.2. *Розглядається стохастичне середовище зі скінченим горизонтом T і кількістю дій N. Кожна дія i має бета-розподіл з невідомим мате-*

матичним сподіванням μ_i. Припустимо, що перша дія є оптимальною без втрати загальності. Тоді при використанні жадібної стратегії за алгоритмом 5.1 має місце наступна нерівність:

$$\mathbb{E}\left[L(T)\right] \leqslant C\sum_{i=2}^{N}(\mu_1-\mu_i) + (T-CN)\sum_{i=2}^{N}(\mu_1-\mu_i)\exp\left(-C\left(\mu_1-\mu_i\right)^2\right).$$

Доведення. Згідно протоколу алгоритму 5.1, кожна дія вибирається рівно C разів під час дослідження і $(T-CN)$ разів під час використання з ймовірністю отримання найбільшого середнього, тобто маємо

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}\right] = \\
&= C + (T-CN)\,\mathbb{P}\left(I_{CN+1}=i\right) = \\
&\leqslant C + (T-CN)\,\mathbb{P}\left(\frac{\sum_{t=1}^{CN}\mathbb{1}_{\{I_t=i\}}\xi_t}{\sum_{t=1}^{CN}\mathbb{1}_{\{I_t=i\}}} \geqslant \max_{j\in\{1,\dots,N\}\setminus\{i\}}\frac{\sum_{t=1}^{CN}\mathbb{1}_{\{I_t=j\}}\xi_t}{\sum_{t=1}^{CN}\mathbb{1}_{\{I_t=j\}}}\right) \leqslant \\
&\leqslant C + (T-CN)\,\mathbb{P}\left(\frac{\sum_{t=1}^{CN}\mathbb{1}_{\{I_t=i\}}\xi_t}{\sum_{t=1}^{CN}\mathbb{1}_{\{I_t=i\}}} \geqslant \frac{\sum_{t=1}^{CN}\mathbb{1}_{\{I_t=1\}}\xi_t}{\sum_{t=1}^{CN}\mathbb{1}_{\{I_t=1\}}}\right) = \\
&= C + (T-CN)\,\mathbb{P}\left(\frac{\sum_{t=1}^{CN}\mathbb{1}_{\{I_t=i\}}\xi_t}{\sum_{t=1}^{CN}\mathbb{1}_{\{I_t=i\}}} - \mu_i - \left(\frac{\sum_{t=1}^{CN}\mathbb{1}_{\{I_t=i\}}\xi_t}{\sum_{t=1}^{CN}\mathbb{1}_{\{I_t=i\}}} - \mu_1\right) \geqslant \mu_1-\mu_i\right) = \\
&= C + (T-CN)\,\mathbb{P}\left(\frac{\sum_{t=1}^{CN}\mathbb{1}_{\{I_t=i\}}\left(\xi_t-\mu_i\right)}{\sum_{t=1}^{CN}\mathbb{1}_{\{I_t=i\}}} - \frac{\sum_{t=1}^{CN}\mathbb{1}_{\{I_t=1\}}\left(\xi_t-\mu_1\right)}{\sum_{t=1}^{CN}\mathbb{1}_{\{I_t=1\}}} \geqslant \mu_1-\mu_i\right),
\end{aligned} \tag{5.3}$$

де $\xi_t-\mu_{I_t}$ є центрованою випадковою величиною з бета-розподілом, згідно з лемою 2.3 отримаємо наступну оцінку ймовірності з (5.3):

$$\begin{aligned}
&\mathbb{P}\left(\frac{\sum_{t=1}^{CN}\mathbb{1}_{\{I_t=i\}}\left(\xi_t-\mu_i\right)}{\sum_{t=1}^{CN}\mathbb{1}_{\{I_t=i\}}} - \frac{\sum_{t=1}^{CN}\mathbb{1}_{\{I_t=1\}}\left(\xi_t-\mu_1\right)}{\sum_{t=1}^{CN}\mathbb{1}_{\{I_t=1\}}} \geqslant \mu_1-\mu_i\right) \leqslant \\
&\quad\leqslant \exp\left(-\frac{\left(\mu_1-\mu_i\right)^2}{2\left(1/\sqrt{CN}\right)^2}\right) = \exp\left(-C\left(\mu_1-\mu_i\right)^2\right).
\end{aligned}$$

Звідси разом з (5.3) та використовуючи лему 2.1 маємо

$$\mathbb{E}\left[L(T)\right] = \sum_{i=1}^{N} \left(\mu_1 - \mu_i\right) \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{I_t = i\}}\right] \leqslant$$
$$\leqslant \sum_{i=1}^{N} (\mu_1 - \mu_i) \left(C + (T - CN) \exp\left(-C\left(\mu_1 - \mu_i\right)^2\right)\right),$$

що і треба було довести. □

5.4. Чисельні експерименти

У цьому підрозділі представлені результати емпіричних тестів для жадібної стратегії. Для цього було розроблено програмне забезпечення [28, 29] з імплементацією алгоритму 5.1 для середовища класу $\mathcal{V}^{\text{Beta}}$ з означення 2.2. Більш детальний опис математичного моделювання надається у розділі 6.

Мета експериментів — показати властивості та асимптотичну поведінку наведеної стратегії у середовищах з різними параметрами. Результати всіх експериментів агреговані з 10000 незалежних тестів і зображені на рисунку 5.1. На графіку показані результати використання жадібної стратегії з оптимальною кількістю досліджень згідно з (5.2) та верхня границя згідно з теоремою 5.1.

Використані параметри середовищ і додаткова інформація результатів наведені в таблиці Б.1.

З отриманих результатів експериментів можна побачити підтвердження теоретичних результатів з теореми 5.1 та наслідку 5.1. Теоретична верхня границя досить близько апроксимує емпіричні результати.

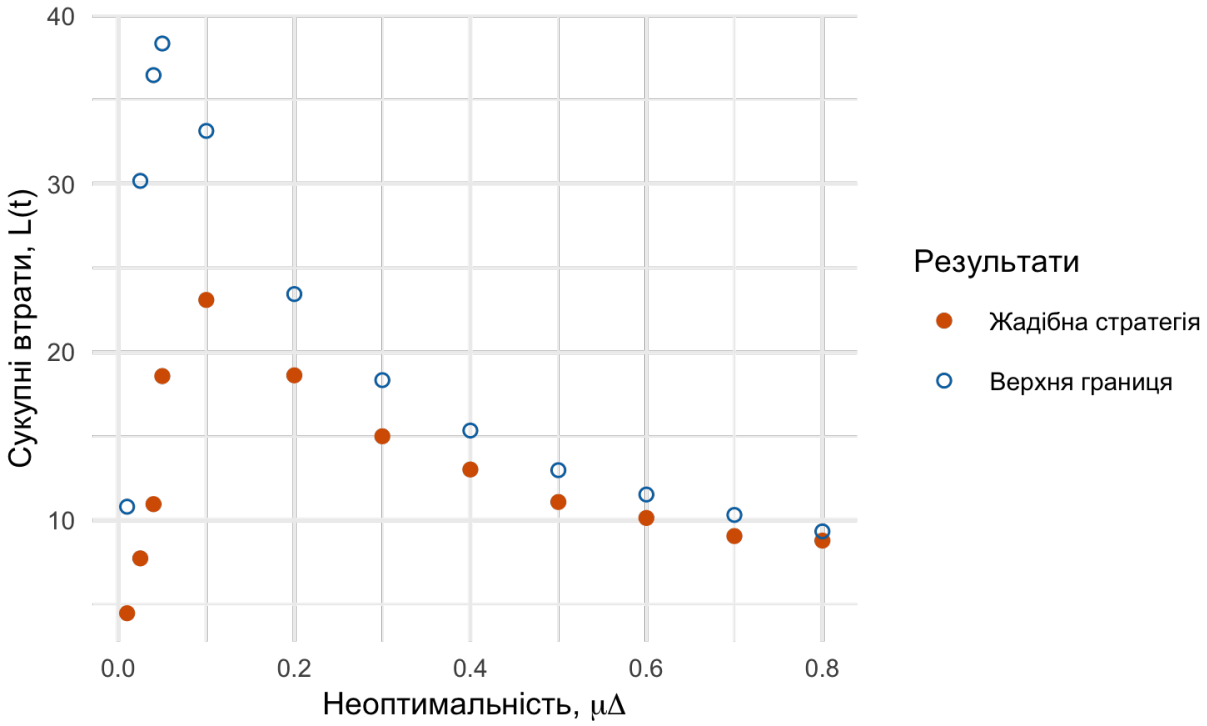


Рис. 5.1: Результати експериментів у стохастичному середовищі зі спостереженнями, які мають бета-розподіл, для жадібної стратегії, які були агреговані з 10000 незалежних тестів. Демонструють залежність сукупних втрат від неоптимальності $\Delta\mu$ у середовищі з двома діями. Результати підтверджують теоретичні висновки щодо отриманої верхньої границі в теоремі 5.1 при використанні оптимізації кількості дослідженнь згідно з (5.2).

Висновки до розділу 5

В цьому розділі була досліджена жадібна стратегія для випадку середовища зі спостереженнями, які мають бета-розподіл. Зокрема, були отримані наступні результати:

- Отримано асимптотичну оцінку очікуваних сукупних втрат з залежністю від неоптимальності дій.
- Отримано оцінку ефективності стратегії для випадку з двома діями та оптимальним вибором кількості досліджень простору варіантів.
- Отримано оцінку ефективності стратегії у загальному випадку.
- Проведено чисельні експерименти, які підтверджують отримані теоретичні результати.

Недоліком жадібної стратегії в порівнянні з баєсовою та стратегією на базі надійного інтервалу є те, що ймовірність отримання лінійних сукупних втрат значно більше, якщо вибір кількості досліджень простору варіантів не є оптимальним. На заданому кроці ця стратегія зовсім припиняє дослідження, коли інші стратегії продовжують в певному обсязі в залежності від межі похибки.

Перевагою стратегії є те, що алгоритм 5.1 не залежить від розподілу середовища.

РОЗДІЛ 6
МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ ТА ПОРІВНЯННЯ СТРАТЕГІЙ

Даний розділ присвячений опису математичного моделювання, яке використовується для чисельних експериментів у попередніх розділах дисертації. Також наведено результати чисельних експериментів, у яких порівнюються розглянуті стратегії у середовищі зі спостереженнями, які мають бета-розподіл:

- стратегія на базі надійного інтервалу за алгоритмом 3.1;
- баєсова стратегія за алгоритмом 4.2;
- жадібна стратегія за алгоритмом 5.1.

6.1. Опис імплементації програмного забезпечення

Наведемо головні реалізовані модулі програмного забезпечення для математичного моделювання та опишемо взаємодію основних компонентів. Весь програмний код математичного моделювання, обробки результатів та побудови графіків опубліковано як ресурс з відкритим кодом [28, 29]. Протягом роботи використовувалися мови програмування Python [74] та R [64].

Для початку нам необхідно створити загальні моделі для середовищ і стратегій з метою подальшого перевикористання при реєстрації різних класів та уніфікації взаємодії між середовищем та стратегією.

Реалізація інтерфейсів середовищ наведена в лістингу 6.1. Використовується імплементація моделі багаторукого бандита з Бернуллі- та бета-розподілом.

Лістинг 6.1: Стохастичне середовище

```
class BetaBandit(Environment):
    def __init__(self, alphas: np.ndarray, betas: np.ndarray, horizon: int):
        self.alphas = alphas
        self.betas = betas

    def draw(self, t: int, action: int) -> float:
        return np.random.beta(a=self.alphas[action], b=self.betas[action])

class BernoulliBandit(Environment):
    def __init__(self, means: np.ndarray, horizon: int):
        self.means = means

    def draw(self, t: int, action: int) -> float:
        return np.random.binomial(n=1, p=self.means[action])
```

В функції draw виконується відбір вибірки за допомоги бібліотеки NumPy [5], де використовується генератор псевдовипадкових чисел PCG-64 описаний в роботі [60]. Для розподілу Бернуллі ми використовуємо Біноміальний розподіл з параметрами ймовірності успіху p та кількості випробувань $n = 1$.

В лістингу 6.2 наведена реалізація стратегії на базі надійного інтервалу за алгоритмом 3.1. Функція select використовується для вибору наступної дії за допомоги функції index, де реалізовано індекс $U_i(t)$ з (3.3).

Лістинг 6.2: Стратегія на базі надійного інтервалу

```
class UCB(Policy):
    def __init__(self, num_actions: int, horizon: int):
        self.num_actions = num_actions
        self.horizon = horizon
        self.cumulative_rewards = np.zeros(num_actions)
        self.actions_count = np.zeros(num_actions)

    def select(self, t: int) -> int:
        if t < self.num_actions:
            return t
        actions = np.arange(self.num_actions)
```

```
12          indexes = np.array([self.index(t=t, action=i) for i in actions])
13          idx = np.where(indexes == np.max(indexes))
14          best_actions = actions[idx]
15          return np.random.choice(best_actions)
16
17      def update(self, t: int, action: int, reward: float):
18          self.cumulative_rewards[action] += reward
19          self.actions_count[action] += 1
20          return
21
22      def index(self, t, action) -> float:
23          mean = self.cumulative_rewards[action] / self.actions_count[action]
24          confidence_radius = math.sqrt(math.log(self.horizon) / (2 * self.
                actions_count[action]))
25          return mean + confidence_radius
```

В лістингу 6.3 наведена реалізація баєсової стратегії за алгоритмом 4.2. Функція select виконує відбір вибірки з кожного розподілу дії та використовується для вибору наступної дії. Функція update виводить новий апостеріорний розподіл за допомогою ієрархічної моделі з (4.2).

Лістинг 6.3: Баєсова стратегія

```
1  class Bayesian(Policy):
2      def __init__(self, name, num_actions: int, horizon: int):
3          self.num_actions = num_actions
4          self.horizon = horizon
5          self.alpha_0 = 1.0
6          self.beta_0 = 1.0
7          self.alphas = np.zeros(num_actions)
8          self.betas = np.zeros(num_actions)
9          self.actions_count = np.zeros(num_actions)
10
11     def select(self, t: int) -> int:
12         samples = np.zeros(self.num_actions)
13         for i in range(self.num_actions):
14             alpha = self.alphas[i] + self.alpha_0
15             beta = self.betas[i] + self.beta_0
16             samples[i] = np.random.beta(alpha, beta)
17         return int(np.argmax(samples))
```

```
18
19      def update(self, t: int, action: int, reward: float):
20          reward = np.random.binomial(n = 1, p = reward)
21          self.alphas[action] = self.alphas[action] + reward
22          self.betas[action] = self.betas[action] + (1.0 - reward)
23          self.actions_count[action] += 1
```

В лістингу 6.4 наведена реалізація жадібної стратегії за алгоритмом 5.1. Функції select та update виконують фіксовану кількість досліджень кожної дії в першій фазі дослідження простору варіантів і роблять вибір емпірично кращої дії для використання у другій фазі. Кількість досліджень контролюється через параметр num_explorations.

Лістинг 6.4: Жадібна стратегія

```
1  class ExploreFirst(Policy):
2      def __init__(self, num_actions: int, horizon: int, num_explorations: int):
3          self.num_actions = num_actions
4          self.num_explorations = num_explorations
5          self.exploration_cumulative_rewards = np.zeros(num_actions)
6          self.best_action = None
7
8      def select(self, t: int) -> int:
9          if t < self.num_explorations * self.num_actions:
10             return t % self.num_actions
11         if self.best_action is None:
12             self.best_action = int(np.argmax(self.
                   exploration_cumulative_rewards))
13         return self.best_action
14
15     def update(self, t: int, action: int, reward: float):
16         if t < self.num_explorations * self.num_actions:
17             self.exploration_cumulative_rewards[action] += reward
```

Маємо наступний алгоритм взаємодії між будь-якою стратегією та будь-яким середовищем за допомогою розглянутих модулів.

Алгоритм 6.1. *Взаємодія між стратегією та середовищем* на скінченому горизонті T.

Крок 1. Покласти $t = 1$.

Крок 2. Призначити наступну дію відповідно до стратегії:

$$I_t = \mathrm{select}(\mathrm{t} = t).$$

Крок 3. Виконати відбір ξ_t з розподілу, пов'язаного з дією I_t:

$$\xi_t = \mathrm{draw}(\mathrm{t} = t, \mathrm{action} = I_t).$$

Крок 4. Оновити отримані емпіричні результати (відповідно до стратегії: індекс на базі надійного інтервалу, апостеріорний розподіл чи вибіркове середнє):

$$\mathrm{update}(\mathrm{t} = t, \mathrm{reward} = \xi_t).$$

Крок 5. Якщо $t > T$, то закінчити виконання алгоритму. Інакше — збільшити t на 1 та перейти до кроку 2.

Таким чином, маємо реалізацію основних компонентів математичного моделювання та алгоритм для отримання результатів, який має бути прозорим. Імплементацію взаємодії, обробки результатів, побудови графіків та інших модулів дивись в роботах опублікованих як ресурс з відкритим кодом [28, 29].

6.2. Чисельні експерименти

У цьому підрозділі представлені результати емпіричних тестів для розглянутих стратегій у стохастичному середовищі зі спостереженнями, які мають бета-розподіл. Результати всіх експериментів агреговані з 10000 незалежних тестів і зображені на рисунку 6.1.

Спочатку на рисунку 6.1a порівнюються асимптотично оптимальні стратегії за означенням 2.10 в середовищах з різною кількістю дій, але однаковим математичним сподіванням оптимальної дії:

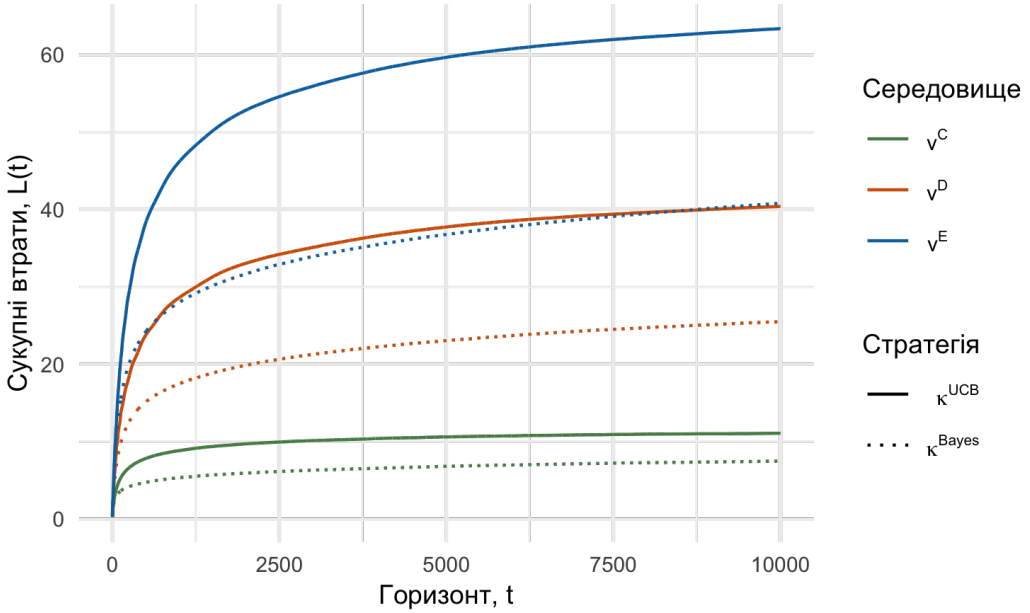
- стратегія на базі надійного інтервалу за алгоритмом 3.1, використовується означення κ^{UCB};
- баєсова стратегія за алгоритмом 4.2, використовується означення κ^{Bayes}.

На графіку можна побачити, що баєсова стратегія має значно менші сукупні втрати в порівняні зі стратегією на базі надійного інтервалу. Це може бути обумовлено впливом вибору значення вірогідного рівня, який було покладено $\delta = 1/T^2$ в алгоритмі 3.1 для стратегії на базі надійного інтервалу. Цей параметр відповідає за темп дослідження простору варіантів і може значно впливати на ситуацію. Знаходження оптимального значення може бути наступним кроком в дослідженні. Використані параметри середовищ наведені в таблиці 3.1.

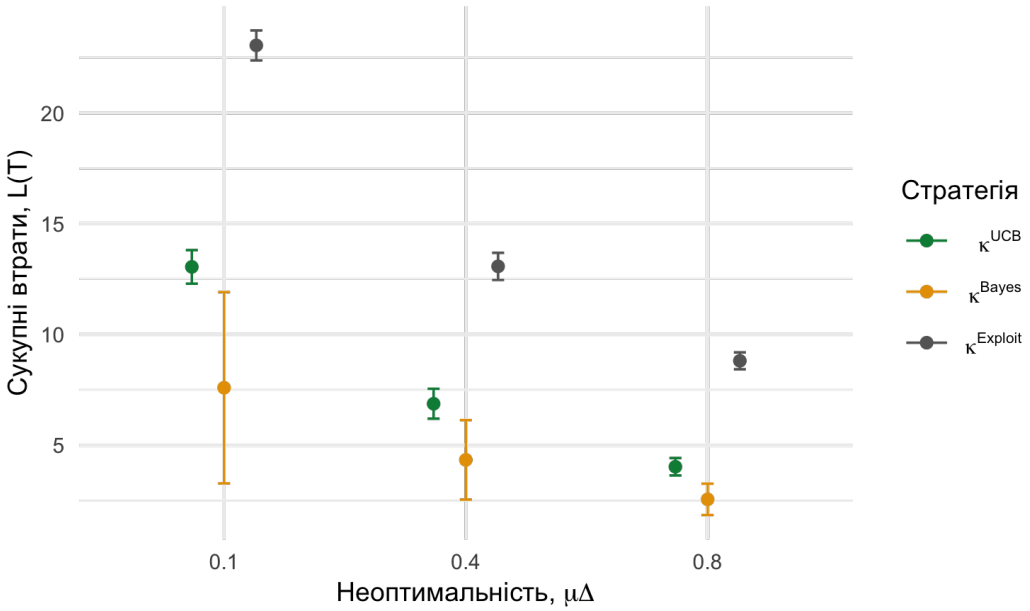
На наступному рисунку 6.1б розглядається середовище з двома діями та різною неоптимальністю $\Delta\mu = |\mu_1 - \mu_2|$ для наступних імплементацій стратегій:

- жадібна стратегія за алгоритмом 3.1 та вибором оптимальної кількості дослідження простору варіантів відповідно до (5.2), використовується означення κ^{Exploit};
- стратегія на базі надійного інтервалу за алгоритмом 3.1, використовується означення κ^{UCB};
- баєсова стратегія за алгоритмом 4.2, використовується означення κ^{Bayes}.

На графіку можна побачити, що баєсова стратегія все ще має значно менші сукупні втрати, але значно більше вибіркове стандартне відхилення. Для покращення ситуації з відхиленням, можна дослідити вибір початкового апріорного розподілу.



(а) Результати експериментів з однаковим математичним сподіванням оптимальної дії та в залежності від кількості дій: (C) дві, (D) чотири та (E) шість



(б) Представлення мінливості даних, один вус представляє одне стандартне відхилення

Рис. 6.1: Результати експериментів у середовищі зі спостереженнями, які мають бета-розподіл, для розглянутих стратегій: на базі надійного інтервалу (κ^{UCB}), баєсової (κ^{Bayes}) та жадібної (κ^{Exploit}).

Висновки до розділу 6

У цьому розділі був наданий детальний опис математичного моделювання та алгоритм взаємодії між агентом та середовищем для вивчення ефективності стратегій, які були розглянуті у попередніх розділах.

Ми порівняли наступні алгоритми у середовищах з різною кількістю дій та неоптимальністю:

- стратегія на базі надійного інтервалу за алгоритмом 3.1;
- баєсова стратегія за алгоритмом 4.2;
- жадібна стратегія за алгоритмом 3.1.

Наведемо наступні властивості та відмінності цих стратегій:

- Ефективність *стратегії на базі надійного інтервалу* залежить від вибору вірогідного рівня. У математичному моделюванні ця стратегія показала одні з найкращих результатів якщо брати до уваги сукупні втрати та вибіркове відхилення разом.
- *Баєсова стратегія* показала найкращі результати щодо вибіркового середнього сукупних втрат. Також ця стратегія додатково має кілька переваг, а саме: можливість обрання апріорного розподілу, що повинно значно покращити результати, коли є ясність вибору; краща стійкість до затримок винагороди ([18]). У випадку присутності затримок, баєсова стратегія продовжує виконувати відбір вибірки, коли поведінка (вибір наступної дії) стратегії на базі надійного інтервалу є детермінованою.
- *Жадібна стратегія*, яка не є асимптотично оптимальною у загальному випадку, показала найгірші результати. Цей метод має кілька переваг — це доступність і легкість застосування.

РОЗДІЛ 7
СТОХАСТИЧНЕ СЕРЕДОВИЩЕ З ДОДАТКОВОЮ ІНФОРМАЦІЄЮ

У даному розділі розглядається середовище, у якому процес винагороди кожної дії залежить від деякої додаткової інформації. На прикладі моделі клінічного випробування з адаптивними стратегіями, де кожна дія представлена окремим препаратом, що досліджується, на результат може впливати певна інформація щодо суб'єкта випробування як, наприклад, вікова категорія чи цілий набір даних.

Ми пропонуємо нове формулювання проблеми, де додаткова інформація це інше спостереження та представлена результатом послідовного кластерного аналізу з високою компактністю та роздільністю ([24, 7]) (чи багатокласовою класифікацією). У цій дисертації наша увага зосереджена на моделі багаторукого бандита, тому використовуємо один з найпростіших алгоритмів кластерного аналізу в чисельних експериментах та опускаємо деталі його дослідження. Більш детальний розгляд послідовного кластерного аналізу наведено у нашій статті [36].

У підрозділі 7.1 ми наведемо асимптотичний аналіз баєсової стратегії у стаціонарному стохастичному середовищі з додатковою інформацією та спостереженнями, які мають бета-розподіл. Отримаємо асимптотичну оцінку очікуваних сукупних втрат з залежністю від неоптимальності дій та у загальному випадку. У підрозділі 7.2 розглядаються вплив помилкової класифікації та можливість зменшення цього впливу. Ми пропонуємо адаптований алгоритм в підрозділі 7.3. У підрозділі 7.4 наведено чисельний експеримент.

Результати даного розділу опубліковані у статті [36].

7.1. Аналіз баєсової стратегії для середовища з додатковою інформацією

Нехай маємо спостереження випадкової величини $Y_t \in \{1, \dots, K\}$ на кожному кроці $t = 1, 2, \dots, T$, яка представляє додаткову інформацію, що доступна на початку кроку. Агент обирає дію I_t із заданої множини $\{1, 2, \dots, N\}$, у відповідь середовище видає винагороду ξ_t з деякого розподілу $P_{I_t \,|\, Y_t}$, пов'язаного з дією I_t за умови Y_t. Припускаємо, що маємо незалежні однаково розподілені випадкові величини, а математичне сподівання існує і є скінченним:

$$\mu_{Y_t, I_t} \coloneqq \mu(I_t \,|\, Y_t) = \int_{-\infty}^{\infty} x \, dP_{I_t \,|\, Y_t}(x).$$

Тобто вибір дії залежить від історії попередніх виборів і їх результатів:

$$(Y_1, I_1, \xi_1, Y_2, I_2, \xi_2, \dots, Y_{t-1}, I_{t-1}, \xi_{t-1}).$$

Для нашого випадку з кластеризацією додаткової інформації, ми робимо припущення, що множина класів $\{1, \dots, K\}$ є скінченною і незмінною. Додамо означення очікуваних сукупних втрат відповідно.

Означення 7.1. *У стаціонарному стохастичному середовищі з додатковою інформацією очікувані сукупні втрати* $\mathbb{E}[L]$ за T кроків визначаються наступним чином:

$$\mathbb{E}[L(T)] = \mathbb{E}\left[\sum_{y=1}^{K} \max_{i=1,\dots,N} \left(\sum_{t=1}^{T} \mathbb{1}_{\{Y_t = y\}} (\mu_{y,i} - \xi_t)\right)\right].$$

Побудуємо алгоритм, який для кожної додаткової інформації $y \in \{1, \dots, K\}$ використовує окрему баєсову стратегію на базі алгоритму 4.2.

Алгоритм 7.1. *Алгоритм баєсової стратегії для середовища з додатковою інформацією та спостереженнями, які мають бета-розподіл.* Розглядається

стохастичне середовище зі скінченим горизонтом T, кількістю додаткової інформації K та дій N. На початку кроку маємо спостереження випадкової величини $Y_t \in \{1, \dots, K\}$. Кожна дія $i \in \{1, \dots, N\}$ має бета-розподіл з невідомим математичним сподіванням за умови Y_t. Вибираючи дію I_t, модель виконує відбір ξ_t з розподілу, пов'язаного з дією I_t за умови Y_t та, як результат, реалізація вибірки стає доступною для стратегії.

Крок 1. Покласти $t = 1$.

Крок 2. Для кожної дії $i \in \{1, \dots, N\}$ та $y \in \{1, \dots, K\}$ покласти

$$\alpha_{y,i} = 1, \ \beta_{y,i} = 1.$$

Крок 3. Для кожної дії $i \in \{1, \dots, N\}$ виконати відбір

$$\hat{\theta}_{Y_t,i} \sim \text{Beta}(\alpha_{Y_t,i}, \beta_{Y_t,i}).$$

Крок 4. Призначити $I_t = \arg\max_{i=1,\dots,N} \hat{\theta}_{Y_t,i}$.

Крок 5. Виконати відбір ξ_t з розподілу, пов'язаного з дією I_t.

Крок 6. Виконати відбір $\eta_t \sim \text{Bern}(\xi_t)$.

Крок 7. Покласти

$$\alpha_{Y_t,I_t} = \alpha_{Y_t,I_t} + \eta_t,$$
$$\beta_{Y_t,I_t} = \beta_{Y_t,I_t} + (1 - \eta_t).$$

Крок 8. Якщо $t > T$, то закінчити виконання алгоритму. Інакше — збільшити t на 1 та перейти до кроку 3.

Отримаємо оцінки сукупних втрат для загального випадку та з залежністю від неоптимальності дій.

Теорема 7.1. *Розглядається стохастичне середовище зі скінченим горизонтом T, кількістю дій N та додатковою інформацією K. Кожна дія $i \in \{1, \dots, N\}$ має бета-розподіл з невідомим математичним сподіванням $\mu_{y,i}$*

за умови інформації $y \in \{1, \dots, K\}$. Тоді при використанні баєсової стратегії за алгоритмом 7.1 має місце наступна нерівність:

$$\mathbb{E}\left[L(T)\right] \leqslant 14\sqrt{KNT}.$$

Доведення. Використаємо результати (4.1) для баєсової стратегії у середовищі без додаткової інформації з означенням сукупних втрат 2.6 через приведення до означення втрат у середовищі з додатковою інформацією 7.1. Зазначимо, що кількість використаних кроків (горизонт) для кожної додаткової інформації y є $\sum_{t=1}^{T} \mathbb{1}_{\{Y_t = y\}}$. Отже, маємо

$$\begin{aligned}
\mathbb{E}\left[L(T)\right] &= \sum_{t=1}^{T} \max_{i=1,\dots,N} \mu_{Y_t,i} - \mathbb{E}\left[\sum_{t=1}^{T} \xi_t\right] = \\
&= \sum_{y=1}^{K} \mathbb{E}\left[\max_{i=1,\dots,N} \sum_{t=1}^{T} \mathbb{1}_{\{Y_t = y\}} \left(\mu_{Y_t,i} - \xi_t\right)\right] = \\
&= \sum_{y=1}^{K} \left(\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{Y_t = y\}}\right] \max_{i=1,\dots,N} \mu_{y,i} - \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{Y_t = y\}} \xi_t\right]\right) \leqslant \\
&\leqslant \sum_{y=1}^{K} 14\sqrt{N\,\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{Y_t = y\}}\right]} = \\
&= 14\,\mathbb{E}\left[\sum_{y=1}^{K} \sqrt{N \sum_{t=1}^{T} \mathbb{1}_{\{Y_t = y\}}}\right]. \qquad (7.1)
\end{aligned}$$

Для того, щоб виразити оцінку через горизонт T замість останньої випадковості $\sum_{t=1}^{T} \mathbb{1}_{\{Y_t = y\}}$, скористаємось тим, що функція $\sqrt{x}$ є опукла вгору. Через нерівність Єнсена отримаємо

$$\sum_{y=1}^{K} \sqrt{N \sum_{t=1}^{T} \mathbb{1}_{\{Y_t = y\}}} = \sqrt{N} K \frac{\sum_{y=1}^{K} \sqrt{\sum_{t=1}^{T} \mathbb{1}_{\{Y_t = y\}}}}{K} \leqslant \sqrt{KNT},$$

та підставимо в (7.1). Тепер маємо

$$\mathbb{E}\left[L(T)\right] \leqslant 14\,\mathbb{E}\left[\sum_{y=1}^{K} \sqrt{N \sum_{t=1}^{T} \mathbb{1}_{\{Y_t = y\}}}\right] \leqslant 14\,\mathbb{E}\left[\sqrt{KNT}\right] = 14\sqrt{KNT}.$$

Що і треба було довести. □

Теорема 7.2. *Розглядається стохастичне середовище зі скінченим горизонтом T, кількістю дій N та додатковою інформацією K. Кожна дія $i \in \{1, \dots, N\}$ має бета-розподіл з невідомим математичним сподіванням $\mu_{y,i}$ за умови додаткової інформації $y \in \{1, \dots, K\}$. Припустимо, що перша дія є оптимальною без втрати загальності. Візьмемо рівномірний розподіл для апріорного розподілу Π_0. Тоді при використанні баєсової стратегії за алгоритмом 7.1 маємо наступну оцінку верхньої границі втрат з залежністю від неоптимальності дій:*

$$\mathbb{E}\left[L(T)\right] \leqslant \max_{y=1,\dots,K} \left(\sum_{i=2}^{N} \frac{1}{\left(\mu_{y,1} - \mu_{y,i}\right)^2} \right) K \log\left(\frac{T}{K}\right).$$

Доведення. Використаємо результати наслідку 4.1 для баєсової стратегії у середовищі без додаткової інформації з означенням сукупних втрат з леми 2.1 через приведення до означення втрат у середовищі з додатковою інформацією. Як і в доведенні попередньої теореми зазначимо, що кількість використаних кроків (горизонт) для кожної додаткової інформації y є $\sum_{t=1}^{T} \mathbb{1}_{\{Y_t = y\}}$. Отже, маємо

$$\begin{aligned}
\mathbb{E}\left[L(T)\right] &= \sum_{t=1}^{T} \max_{i=1,\dots,N} \mu_{Y_t,i} - \mathbb{E}\left[\sum_{t=1}^{T} \xi_t\right] = \\
&= \sum_{y=1}^{K} \mathbb{E}\left[\max_{i=1,\dots,N} \sum_{t=1}^{T} \mathbb{1}_{\{Y_t = y\}} \left(\mu_{Y_t,i} - \xi_t\right)\right] = \\
&= \sum_{y=1}^{K} \left(\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{Y_t = y\}}\right] \max_{i=1,\dots,N} \mu_{y,i} - \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{Y_t = y\}} \xi_t\right] \right) \leqslant \\
&\leqslant \sum_{y=1}^{K} \left(\sum_{i=1}^{N} \frac{1}{\left(\mu_{y,1} - \mu_{y,i}\right)^2} \right)^2 \log\left(\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\{Y_t = y\}}\right] \right) \leqslant \\
&\leqslant \max_{y=1,\dots,K} \left(\sum_{i=1}^{N} \frac{1}{\left(\mu_{y,1} - \mu_{y,i}\right)^2} \right)^2 \mathbb{E}\left[\sum_{y=1}^{K} \log\left(\sum_{t=1}^{T} \mathbb{1}_{\{Y_t = y\}}\right)\right]. \qquad (7.2)
\end{aligned}$$

Для того, щоб виразити сукупні втрати через горизонт T, оцінимо

$$\sum_{y=1}^{K} \log \left(\sum_{t=1}^{T} \mathbb{1}_{\{Y_t = y\}} \right)$$

за допомогою того, що горизонт можна виразити через

$$T = \sum_{y=1}^{K} \sum_{t=1}^{T} \mathbb{1}_{\{Y_t = y\}}$$

та того, що максимальний добуток

$$\prod_{y=1}^{K} \sum_{t=1}^{T} \mathbb{1}_{\{Y_t = y\}}$$

виникає при рівних доданках. Отже, маємо

$$\sum_{y=1}^{K} \log \left(\sum_{t=1}^{T} \mathbb{1}_{\{Y_t = y\}} \right) = \log \left(\prod_{y=1}^{K} \sum_{t=1}^{T} \mathbb{1}_{\{Y_t = y\}} \right) \leqslant \log \left(\frac{T}{K} \right)^K .$$

Підставимо цю оцінку в (7.2) та отримаємо

$$\begin{aligned} \mathbb{E}\left[L(T)\right] &\leqslant \max_{y=1,\dots,K} \left(\sum_{i=1}^{N} \frac{1}{(\mu_{y,1} - \mu_{y,i})^2} \right)^2 \mathbb{E}\left[\sum_{y=1}^{K} \log \left(\sum_{t=1}^{T} \mathbb{1}_{\{Y_t = y\}} \right) \right] \leqslant \\ &\leqslant \max_{y=1,\dots,K} \left(\sum_{i=1}^{N} \frac{1}{(\mu_{y,1} - \mu_{y,i})^2} \right)^2 \log \left(\frac{T}{K} \right)^K . \end{aligned}$$

Що доводить теорему. □

7.2. Аналіз впливу помилкової класифікації

В продовження теми середовища з додатковою інформацією у цьому підрозділі ми припускаємо, що доступне для стратегії спостереження $\hat{Y}_t$ є результатом послідовного кластерного аналізу з кількістю класів K. Справжнє спостереження додаткової інформації Y_t, яке породжує множину розподілів $(P_{Y_t,i} : i \in \{1, \dots, N\})$ на кроці t, не є доступним для стратегії. Таким чином вибір дії залежить від історії спостережень класифікації

поточної додаткової інформації, попередніх виборів та їх результатів:

$$(\hat{Y}_1, I_1, \xi_1, \hat{Y}_2, I_2, \xi_2, \ldots, \hat{Y}_{t-1}, I_{t-1}, \xi_{t-1}).$$

Розглянемо вплив помилкової класифікації на стратегію. У випадку $\hat{Y}_t \neq Y_t$ при значній різниці в математичному сподіванні $\mu_{\hat{Y}_t,A_t} \gg \mu_{Y_t,A_t}$ ($\mu_{\hat{Y}_t,A_t} \ll \mu_{Y_t,A_t}$) ми будемо недооцінювати (переоцінювати) параметр розподілу $P_{\hat{Y}_t,A_t}$ використовуючи вибірку з P_{Y_t,A_t}.

Зробимо припущення, що у процесі винагороди має місце ґауссівський шум. Будемо використовувати наступну ієрархічну модель:

$$\eta_t \sim \text{Bern}\left(\theta'_{\hat{Y}_t,A_t}\right),$$
$$\theta'_{\hat{Y}_t,A_t} \sim \mathcal{N}(\mu_{\hat{Y}_t,A_t}, \sigma_t^2, a = 0, b = 1),$$

де $\mathcal{N}(\mu, \sigma_t^2, a, b)$ — усічений нормальний розподіл з щільністю

$$f(x; \mu, \sigma^2, a, b) = \frac{1}{\sigma} \frac{\varphi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)},$$

у якому $\varphi(\cdot)$ та $\Phi(\cdot)$ — щільність та функція розподілу стандартного нормального розподілу відповідно. Знайдемо щільність розподілу випадкової величини η_t за умови $\theta'_{\hat{Y}_t,A_t}$:

$$f_{\eta_t|\theta'_{\hat{Y}_t,A_t}}(x|p) = \int_0^1 (px + (1-p)(1-x)) \frac{\exp(-(p - \mu_{\hat{Y}_t,A_t})^2/(2\sigma_t^2))}{\sqrt{2\pi}\sigma_t \left(\Phi\left(\frac{1-\mu_{\hat{Y}_t,A_t}}{\sigma_t}\right) - \Phi\left(\frac{-\mu_{\hat{Y}_t,A_t}}{\sigma_t}\right)\right)} \, dp.$$

Зробимо підстановку

$$u = p - \mu_{\hat{Y}_t,A_t}, \; du = dp$$

та виразимо через два інтеграли наступним чином:

$$f_{\eta_t|\theta'_{\hat{Y}_t,A_t}}(x|p) =$$

$$= (\mu_{\hat{Y}_t,A_t}x + (1-\mu_{\hat{Y}_t,A_t})(1-x)) \int\limits_{-\mu_{\hat{Y}_t,A_t}}^{1-\mu_{\hat{Y}_t,A_t}} \frac{\exp(-u^2/(2\sigma_t^2))}{\sqrt{2\pi}\sigma_t\left(\Phi\left(\frac{1-\mu_{\hat{Y}_t,A_t}}{\sigma_t}\right) - \Phi\left(\frac{-\mu_{\hat{Y}_t,A_t}}{\sigma_t}\right)\right)}\,du +$$

$$+ \frac{(2x-1)}{\sqrt{2\pi}\sigma_t\left(\Phi\left(\frac{1-\mu_{\hat{Y}_t,A_t}}{\sigma_t}\right) - \Phi\left(\frac{-\mu_{\hat{Y}_t,A_t}}{\sigma_t}\right)\right)} \int\limits_{-\mu_{\hat{Y}_t,A_t}}^{1-\mu_{\hat{Y}_t,A_t}} \exp(-u^2/(2\sigma_t^2))\,du,$$

що дає щільність усіченого нормального розподілу у першому інтегралі. Обчислимо другий інтеграл:

$$f_{\eta_t|\theta'_{\hat{Y}_t,A_t}}(x|p) = \mu_{\hat{Y}_t,A_t}x + (1-\mu_{\hat{Y}_t,A_t})(1-x) +$$

$$+ (2x-1)\frac{\sigma_t^2\exp(-u^2/(2\sigma_t^2))}{\sqrt{2\pi}\sigma_t\left(\Phi\left(\frac{1-\mu_{\hat{Y}_t,A_t}}{\sigma_t}\right) - \Phi\left(\frac{-\mu_{\hat{Y}_t,A_t}}{\sigma_t}\right)\right)}\Bigg|_{-\mu_{\hat{Y}_t,A_t}}^{1-\mu_{\hat{Y}_t,A_t}} +$$

$$+ \mu_{\hat{Y}_t,A_t}x + (1-\mu_{\hat{Y}_t,A_t})(1-x) +$$

$$+ (2x-1)\frac{\sigma_t\left(\exp(-\mu_{\hat{Y}_t,A_t}^2/(2\sigma_t^2)) - \exp(-(1-\mu_{\hat{Y}_t,A_t})^2/(2\sigma_t^2))\right)}{\sqrt{2\pi}\left(\Phi\left(\frac{1-\mu_{\hat{Y}_t,A_t}}{\sigma_t}\right) - \Phi\left(\frac{-\mu_{\hat{Y}_t,A_t}}{\sigma_t}\right)\right)}.$$

Таким чином, приходимо до наступної щільності розподілу випадкової величини η_t:

$$f_{\eta_t}(x) = \begin{cases} (1-\mu_{\hat{Y}_t,A_t}) - \frac{\sigma_t\left(\exp(-\mu_{\hat{Y}_t,A_t}^2/(2\sigma_t^2)) - \exp(-(1-\mu_{\hat{Y}_t,A_t})^2/(2\sigma_t^2))\right)}{\sqrt{2\pi}\left(\Phi\left(\frac{1-\mu_{\hat{Y}_t,A_t}}{\sigma_t}\right) - \Phi\left(\frac{-\mu_{\hat{Y}_t,A_t}}{\sigma_t}\right)\right)} & \text{якщо } x=0, \\ \mu_{\hat{Y}_t,A_t} + \frac{\sigma_t\left(\exp(-\mu_{\hat{Y}_t,A_t}^2/(2\sigma_t^2)) - \exp(-(1-\mu_{\hat{Y}_t,A_t})^2/(2\sigma_t^2))\right)}{\sqrt{2\pi}\left(\Phi\left(\frac{1-\mu_{\hat{Y}_t,A_t}}{\sigma_t}\right) - \Phi\left(\frac{-\mu_{\hat{Y}_t,A_t}}{\sigma_t}\right)\right)} & \text{якщо } x=1, \\ 0 & \text{інакше.} \end{cases}$$

При $\sigma_t \to 0$ отримана функція збігається з щільністю розподілу Бернуллі з параметром $\mu_{\hat{Y}_t,A_t}$. Якби в нас був доступ до σ_t, ми могли б зважити вибірку для зменшення шуму у спосіб, подібний до середнього зваження з оберненої дисперсією ([47]):

$$\frac{\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}\eta_t/\sigma_t^2}{\sum_{t=1}^{T}\mathbb{1}_{\{I_t=i\}}/\sigma_t^2.}$$

Побудуємо алгоритм на базі цих результатів у наступному підрозділі.

7.3. Алгоритм з урахуванням коефіцієнта ймовірності правильної класифікації

Нехай на початку кожного кроку додатково до поточного класу $\hat{Y}_t$ маємо ймовірність правильної класифікації $\lambda_t \in [0, 1]$. Застосуємо метод подібний до середнього зваження з оберненою дисперсією, який зважує вибірку обернено пропорційно до дисперсії (чи прямо пропорційно до влучності $1/\sigma^2$). Для нашого випадку будемо вважати отриману ймовірність правильної класифікації влучністю. Отже, обчислимо параметри баєсової стратегії алгоритму 7.1 для дії i наступним чином:

$$\alpha_{\hat{Y}_t,i} = \alpha_0 + \sum_{s=1}^{t} \mathbb{1}_{\left\{\hat{Y}_t = y\right\}} \mathbb{1}_{\{I_t = i\}} \lambda_s \eta_s,$$

$$\beta_{\hat{Y}_t,i} = \beta_0 + \sum_{s=1}^{t} \mathbb{1}_{\left\{\hat{Y}_t = y\right\}} \mathbb{1}_{\{I_t = i\}} \lambda_s (1 - \eta_s),$$

що використаємо у наступному алгоритмі.

Алгоритм 7.2. *Алгоритм баєсової стратегії з урахуванням коефіцієнта ймовірності правильної класифікації для середовища з додатковою інформацією та спостереженнями, які мають бета-розподіл.* Розглядається стохастичне середовище зі скінченим горизонтом T, кількістю додаткової інформації K та дій N. На початку кроку маємо спостереження випадкових величин $\hat{Y}_t$ та λ_t. Кожна дія $i \in \{1, \ldots, N\}$ має бета-розподіл з невідомим математичним сподіванням за умови Y_t. Вибираючи дію I_t, модель виконує відбір ξ_t з розподілу, пов'язаного з дією I_t за умови Y_t та, як результат, реалізація вибірки стає доступною для стратегії.

Крок 1. Покласти $t = 1$.

Крок 2. Для кожної дії $i \in \{1, \ldots, N\}$ та $y \in \{1, \ldots, K\}$ покласти

$$\alpha_{y,i} = 1, \ \beta_{y,i} = 1.$$

Крок 3. Для кожної дії $i \in \{1, \dots, N\}$ виконати відбір

$$\hat{\theta}_{\hat{Y}_t,i} \sim \text{Beta}(\alpha_{\hat{Y}_t,i}, \beta_{\hat{Y}_t,i}).$$

Крок 4. Призначити $I_t = \arg\max_{i=1,\dots,N} \hat{\theta}_{\hat{Y}_t,i}$.

Крок 5. Виконати відбір ξ_t з розподілу, пов'язаного з дією I_t.

Крок 6. Виконати відбір $\eta_t \sim \text{Bern}(\xi_t)$.

Крок 7. Покласти

$$\alpha_{\hat{Y}_t,I_t} = \alpha_{\hat{Y}_t,I_t} + \lambda_t \eta_t,$$
$$\beta_{\hat{Y}_t,I_t} = \beta_{\hat{Y}_t,I_t} + \lambda_t (1 - \eta_t).$$

Крок 8. Якщо $t > T$, то закінчити виконання алгоритму. Інакше — збільшити t на 1 та перейти до кроку 3.

Проаналізуємо цей алгоритм у найгіршому випадку, коли маємо високу невпевненість в класифікації на кожному кроці.

Теорема 7.3. *Розглядається стохастичне середовище зі скінченим горизонтом T, кількістю дій N та додатковою інформацією K. На початку кроку маємо спостереження випадкових величин $\hat{Y}_t$ та λ_t. Кожна дія $i \in \{1, \dots, N\}$ має бета-розподіл з невідомим математичним сподіванням $\mu_{y,i}$ за умови інформації $y \in \{1, \dots, K\}$. Припустимо, що перша дія є оптимальною без втрати загальності. Візьмемо рівномірний розподіл для апріорного розподілу Π_0. Нехай маємо $\lambda_t = 0$ на кожному кроці t. Тоді при використанні баєсової стратегії за алгоритмом 7.2 маємо наступну оцінку з залежністю від неоптимальності дій:*

$$\mathbb{E}\left[L(T)\right] \leqslant \frac{T}{N} \sum_{i=2}^{N} \max_{y=1,\dots,K} \left(\mu_{y,1} - \mu_{y,i}\right).$$

Доведення. Згідно протоколу алгоритму 7.2 маємо $\alpha_{y,i} = \beta_{y,i} = 1$ для всіх $y \in \{1, \dots, K\}$ та $i \in \{1, \dots, N\}$ на всіх кроках, тобто стандартний

рівномірний розподіл. Таким чином, на кожному кроці t

$$\mathbb{E}\left[\mathbb{1}_{\{I_t=i\}} \mid \hat{Y}_t\right] = \mathbb{E}\left[\mathbb{1}_{\{I_t=i\}} \mid Y_t\right] = \frac{1}{N}. \tag{7.3}$$

Так як для всіх $t \in [T]$ виконується

$$\sum_{y=1}^{K}\sum_{i=1}^{N} \mathbb{1}_{\{Y_t=y\}}\mathbb{1}_{\{I_t=i\}} = 1,$$

то виразимо сукупні втрати наступним чином:

$$\begin{aligned}\mathbb{E}\left[L(T)\right] &= \sum_{t=1}^{T}\mu_{Y_t,1} - \mathbb{E}\left[\sum_{t=1}^{T}\xi_t\right] = \\ &= \sum_{y=1}^{K}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{1}_{\{Y_t=y\}}\mathbb{1}_{\{I_t=i\}}\left(\mu_{Y_t,1}-\xi_t\right)\right].\end{aligned}$$

Отже, використовуючи правило повного математичного сподівання та (7.3) отримаємо наступну оцінку втрат з останньої рівності:

$$\begin{aligned}\mathbb{E}\left[L(T)\right] &= \sum_{y=1}^{K}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{Y_t=y\}}\mathbb{1}_{\{I_t=i\}}\left(\mu_{Y_t,1}-\xi_t\right) \mid Y_t, I_t\right]\right] = \\ &= \sum_{y=1}^{K}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{1}_{\{Y_t=y\}}\mathbb{1}_{\{I_t=i\}}\mathbb{E}\left[\left(\mu_{Y_t,1}-\xi_t\right) \mid Y_t, I_t\right]\right] = \\ &= \sum_{y=1}^{K}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{1}_{\{Y_t=y\}}\mathbb{1}_{\{I_t=i\}}\left(\mu_{Y_t,1}-\mu_{Y_t,I_t}\right)\right] = \\ &= \sum_{y=1}^{K}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{1}_{\{Y_t=y\}}\mathbb{1}_{\{I_t=i\}}\right]\left(\mu_{y,1}-\mu_{y,i}\right) = \\ &= \sum_{y=1}^{K}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{Y_t=y\}}\mathbb{1}_{\{I_t=i\}} \mid Y_t\right]\right]\left(\mu_{y,1}-\mu_{y,i}\right) = \\ &= \sum_{y=1}^{K}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{1}_{\{Y_t=y\}}\mathbb{E}\left[\mathbb{1}_{\{I_t=i\}} \mid Y_t\right]\right]\left(\mu_{y,1}-\mu_{y,i}\right) = \\ &= \sum_{y=1}^{K}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{N}\mathbb{E}\left[\mathbb{1}_{\{Y_t=y\}}\right]\left(\mu_{y,1}-\mu_{y,i}\right) = \end{aligned}$$

$$= \sum_{y=1}^{K} \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{1}{N} \mathbb{E}\left[\mathbb{1}_{\{Y_t = y\}}\right] (\mu_{y,1} - \mu_{y,i}) \leqslant$$

$$\leqslant \frac{1}{N} \sum_{i=1}^{N} \max_{y=1,\ldots,K} (\mu_{y,1} - \mu_{y,i}) \sum_{y=1}^{K} \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}_{\{Y_t = y\}}\right] =$$

$$= \frac{T}{N} \sum_{i=1}^{N} \max_{y=1,\ldots,K} (\mu_{y,1} - \mu_{y,i}) .$$

Що і треба було довести. □

У випадку високої впевненості $\lambda_t = 1$ та відсутності помилкової класифікації на кожному кроці t маємо результати з теореми 7.2:

$$\mathbb{E}[L(T)] \leqslant \max_{y=1,\ldots,K} \left(\sum_{i=2}^{N} \frac{1}{(\mu_{y,1} - \mu_{y,i})^2} \right) K \log\left(\frac{T}{K}\right).$$

7.4. Чисельні експерименти

У даному експерименті ми демонструємо потенціал використання коефіцієнта ймовірності правильної класифікації для баєсової стратегії в середовищі з різним відсотком неправильної класифікації. Маємо декілька етапів алгоритму:

1. Послідовний кластерний аналіз за результатом класифікації поточної додаткової інформації $\hat{Y}_t$. Застосовується метод K-найближчих сусідів ([24, 38]).
2. Отримання коефіцієнта ймовірності правильної класифікації з нечіткого кластерного аналізу ([23]):

$$\lambda_t = \sum_{y=1}^{K} \left(\frac{\|x_t - \hat{c}_{\hat{Y}_t}\|}{\|x_t - \hat{c}_y\|} \right)^{-2/(l-1)} \in [0, 1],$$

   де x_t — об'єкт поточної додаткової інформації, $\hat{c}_y$ — центроїд класу y та l — коефіцієнт нечіткості.
3. Використання баєсової стратегії для середовища з додатковою інформацією.

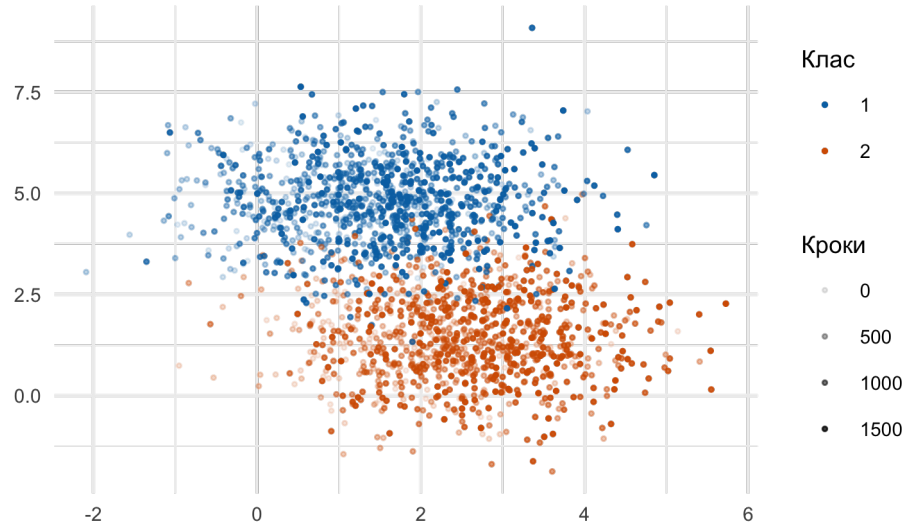
Ми розглядаємо випадки додаткової інформації з двома класами по 10 дій у кожному у вигляді послідовного кластерного аналізу, приклад яких наведено на рисунку 7.1а. Математичні сподівання (параметри розподілів) дій згенеровані випадковим чином. Мета експериментів — порівняти роботу алгоритмів баєсової стратегії без урахування коефіцієнта λ_t 7.1 та з урахуванням за алгоритмом 7.2 в середовищах з різними відсотками правильної класифікації.

Результати всіх експериментів агреговані з 10000 незалежних тестів і зображені на рисунку 7.1б. На цих графіках можна побачити, що баєсова стратегія, яка використовує коефіцієнт ймовірності правильної класифікації, має до 5% менші сукупні втрати.

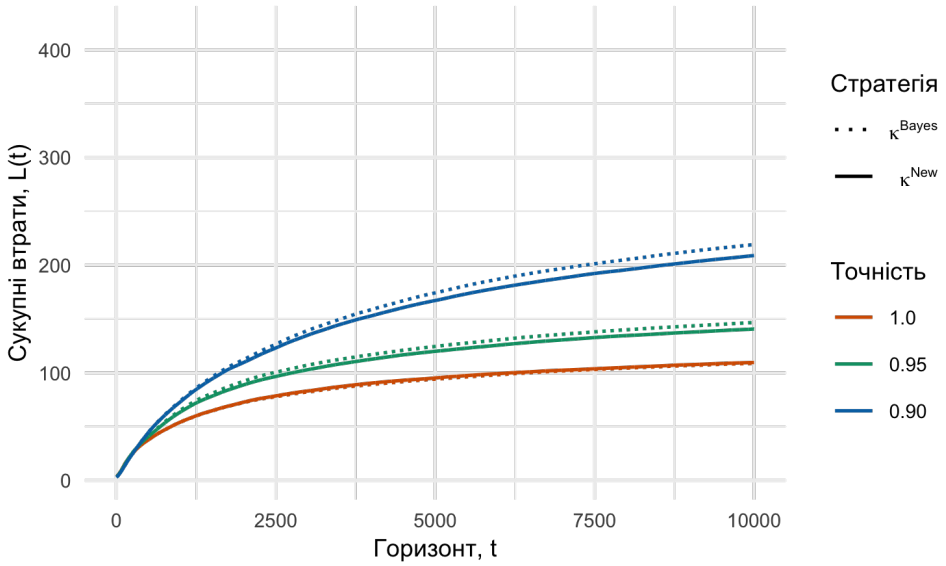
Висновки до розділу 7

У цьому розділі був наданий опис математичної моделі багаторукого бандита у середовищі, у якому процес винагороди кожної дії залежить від деякої додаткової інформації. Зокрема, були отримані наступні результати:

- Наведено асимптотичний аналіз баєсової стратегії у стохастичному середовищі з додатковою інформацією та спостереженнями, які мають бета-розподіл. Отримані оцінки верхньої границі очікуваних сукупних втрат з залежністю від неоптимальності дій та у загальному випадку.
- Розглянуто вплив помилкової класифікації та спосіб зменшення цього впливу. Побудовано новий алгоритм з використанням зважування вибірки для зменшення шуму у спосіб, подібний до середнього зважування з оберненою дисперсією.
- Проведено чисельні експерименти, які підтверджують теоретичні результати.



(а) Додактова інформація у вигляді кластерів



(б) Сукупні втрати на кроках з правильною класифікацією поточної додаткової інформації

Рис. 7.1: Результати експериментів в середовищах з різними відсотками правильної класифікації (точністю) в середньому на горизонті: 100%, 95% та 90%. Використовуються баєсові стратегії за алгоритмом 7.1 (κ^{Bayes}) та стратегія з коефіцієнтом λ_t за алгоритмом 7.2 (κ^{New}).

ВИСНОВКИ

У роботі були розглянуті асимптотичні властивості стратегій у стохастичному середовищі зі спостереженнями, які мають бета-розподіл. Застосовану техніку доведень, запровадженні означення, допоміжні твердження та леми можна використовувати для подальших досліджень в цьому та суміжних напрямках. Серед отриманих результатів дисертаційної роботи слід виділити наступні:

1. Розглянуто стохастичне середовище та наведено означення класів для різних випадків розподілів. Отримано функцію очікуваних сукупних втрат з залежністю від неоптимальності дій у даному середовищі. Зроблено аналіз прикладів неоптимальних стратегій та простих випадків.
2. Адаптовано алгоритм для стратегії на базі надійного інтервалу. Отримана асимптотична оцінка очікуваних сукупних втрат. Показано, що дана стратегія є асимптотично оптимальною для розглянутого випадку середовища.
3. Побудовано алгоритм для баєсової стратегії. Отримано асимптотичну оцінку очікуваних сукупних втрат з залежністю від неоптимальності дій. Отримано баєсову асимптотичну оцінку очікуваних сукупних втрат без залежності від неоптимальності дій та припущень щодо апріорного розподілу. Показано, що баєсова стратегія є асимптотично оптимальною для розглянутого випадку середовища.
4. Отримано оцінку ефективності стратегії для випадку з двома діями та оптимальним вибором кількості досліджень простору варіантів для жадібної стратегії.

5. Наведено асимптотичний аналіз баєсової стратегії у стохастичному середовищі з додатковою інформацією. Отримані оцінки верхньої границі очікуваних сукупних втрат з залежністю від неоптимальності дій та у загальному випадку. Побудовано новий алгоритм з використанням зважування вибірки для зменшення шуму у спосіб, подібний до середнього зважування з оберненою дисперсією.
6. Проведено моделювання запропонованих у роботі алгоритмів стратегій у стохастичному середовищі з різними параметрами розподілів та кількістю дій. Одержані чисельні результати показали, що стратегії є асимптотично оптимальними згідно з отриманими оцінками верхніх границь втрат.
7. Розроблено програмне забезпечення та бібліотеки, які опубліковані як ресурс з відкритим кодом [28, 29]. Використовувалися мови програмування Python [74] та R [64] з додатковими бібліотеками, які знаходяться під ліцензіями BSD-3, PSF і MIT у вільному доступі для наукових досліджень.

Також варто зазначити, що потенціал розроблених стратегій для середовища зі спостереженнями, які мають бета-розподіл, демонструється в одній з опублікованих статей ([34]), де наведено симуляцію експерименту з використанням набору даних, отриманих у результаті реальних клінічних випробувань.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] 1 - The Distribution of the Estimates for the Norm of Sub-Gaussian Stochastic Processes / Kozachenko Y., Pogorilyak O., Rozora I., and Tegza A. // Simulation of Stochastic Processes with Given Accuracy and Reliability / ed. by Kozachenko Y., Pogorilyak O., Rozora I., Tegza A. — Elsevier, 2016. — P. 1–70.

[2] Adaptive designs in clinical trials: why use them, and how to run and report them / Pallmann P., Bedding A. W., Choodari-Oskooei B., Dimairo M., Flight L., Hampson L. V., Holmes J., Mander A. P., Odondi L., Sydes M. R., Villar S. S., Wason J. M. S., Weir C. J., Wheeler G. M., Yap C., and Jaki T. // BMC Medicine. — 2018. — Dec. — Vol. 16, no. 1. — P. 29.

[3] Agrawal S., Goyal N. Analysis of Thompson Sampling for the Multi-Armed Bandit Problem // Conference on learning theory / JMLR Workshop and Conference Proceedings. — 2012. — P. 39–1.

[4] Agrawal S., Goyal N. Near-optimal regret bounds for thompson sampling // Journal of the ACM (JACM). — 2017. — Vol. 64, no. 5. — P. 1–24.

[5] Array programming with NumPy / Harris C. R., Millman K. J., van der Walt S. J., Gommers R., Virtanen P., Cournapeau D., Wieser E., Taylor J., Berg S., Smith N. J., Kern R., Picus M., Hoyer S., van Kerkwijk M. H., Brett M., Haldane A., del Río J. F., Wiebe M., Peterson P., Gérard-Marchant P., Sheppard K., Reddy T., Weckesser W., Abbasi H., Gohlke C., and Oliphant T. E. // Nature. — 2020. — Sep. — Vol. 585, no. 7825. — P. 357–362.

[6] Arrow K. J., Blackwell D., Girshick M. A. Bayes and minimax solutions of sequential decision problems // Econometrica, Journal of the Econometric Society. — 1949. — P. 213–244.

[7] Ashour W., Fyfe C. Online clustering algorithms // International journal of neural systems. — 2008. — Vol. 18. — P. 185–94.

[8] Auer P., Cesa-Bianchi N., Fischer P. Finite-Time Analysis of the Multiarmed Bandit Problem // Machine Learning. — 2002. — Vol. 47, no. 2. — P. 235–256.

[9] Awerbuch B., Kleinberg R. D. Adaptive routing with end-to-end feedback: distributed learning and geometric approaches // Proceedings of the thirty-sixth annual ACM symposium on Theory of computing - STOC '04. — Chicago, IL, USA : ACM Press. — 2004. — P. 45.

[10] Bayesian reinforcement learning: A survey / Ghavamzadeh M., Mannor S., Pineau J., Tamar A., et al. // Foundations and Trends® in Machine Learning. — 2015. — Vol. 8, no. 5-6. — P. 359–483.

[11] Bellman R. Dynamic programming // Science. — 1966. — Vol. 153, no. 3731. — P. 34–37.

[12] Berry D. A. A Bernoulli two-armed bandit // The Annals of Mathematical Statistics. — 1972. — P. 871–897.

[13] Bretagnolle J., Huber C. Estimation des densités : Risque minimax // Séminaire de Probabilités XII / ed. by Dellacherie C., Meyer P. A., Weil M. — Berlin, Heidelberg : Springer Berlin Heidelberg. — 1978. — P. 342–363.

[14] Bubeck S. Regret analysis of stochastic and nonstochastic multi-armed bandit problems // Foundations and Trends in Machine Learning. — 2012. — Vol. 5, no. 1. — P. 1–122.

[15] Bubeck S., Liu C.-Y. Prior-free and prior-dependent regret bounds for thompson sampling // Advances in neural information processing systems. — 2013. — Vol. 26.

[16] Buldygin V. V., Kozachenko Y. V. Metric Characterization of Random Variables and Random Processes. — American Mathematical Soc., 2000. — Vol. 188.

[17] Burnetas A. N., Katehakis M. N. Optimal adaptive policies for sequential allocation problems // Advances in Applied Mathematics. — 1996. — Vol. 17,

no. 2. — P. 122–142.

[18] Chapelle O., Li L. An empirical evaluation of thompson sampling // Advances in neural information processing systems. — 2011. — Vol. 24.

[19] Chernoff H. A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations // The Annals of Mathematical Statistics. — 1952. — Vol. 23, no. 4. — P. 493 – 507.

[20] Chow S.-C., Chang M. Adaptive design methods in clinical trials – a review // Orphanet Journal of Rare Diseases. — 2008. — Dec. — Vol. 3, no. 1. — P. 11.

[21] A contextual-bandit approach to personalized news article recommendation / Li L., Chu W., Langford J., and Schapire R. E. // Proceedings of the 19th international conference on World wide web - WWW '10. — Raleigh, North Carolina, USA : ACM Press. — 2010. — P. 661.

[22] Den Boer A. V. Dynamic pricing and learning: historical origins, current research, and new directions // Surveys in operations research and management science. — 2015. — Vol. 20, no. 1. — P. 1–18.

[23] Duda R. O., Hart P. E., Stork D. G. Fuzzy k-Means Clustering // Pattern Classification 2nd Edition. — John Wiley & Sons, 2001. — P. 528–530.

[24] Duda R. O., Hart P. E., Stork D. G. On-line clustering // Pattern Classification 2nd Edition. — John Wiley & Sons, 2001. — P. 559–565.

[25] Dzhoha A. Модель багаторукого бандита у стохастичному середовищі та чисельні експерименти // VIII Всеукраїнська науково-практична конференція «Інформаційні технології — 2021. Математичне моделювання та обчислювальні методи». — 2021. — С. 176–177.

[26] Dzhoha A. Bernoulli multi-armed bandit problem under delayed feedback // Bulletin of Taras Shevchenko National University of Kyiv. Series: Physics and Mathematics. — 2021. — no. 1. — P. 20–26.

[27] Dzhoha A. Sequential resource allocation in a stochastic environment: an overview and numerical experiments // Bulletin of Taras Shevchenko Na-

tional University of Kyiv. Series: Physics and Mathematics. — 2021. — no. 3. — P. 13–25.

[28] Dzhoha A. Multi-armed bandit problem under delayed deedback: numerical experiments. — https://github.com/djo/delayed-bandit/. — 2022.

[29] Dzhoha A. Contextual multi-armed bandit problem with online clustering: numerical experiments. — https://github.com/djo/bandit-with-online-clustering. — 2023.

[30] Dzhoha A., Lebedev E. Sequential resource allocation under multi-armed bandit model with delays // XXXVI International Conference «Problems of Decision Making under Uncertainties». — 2021. — P. 37–38. — Access mode: http://www.pdmu.univ.kiev.ua/PDMU2021/PDMU2021Skhidnytsia.pdf.

[31] Dzhoha A., Rozora I. Multi-armed bandit problem with online clustering as side information // Abstracts of 8th International Congress of Computational Engineering and Sciences ESCO. — 2022. — June. — Access mode: https://www.esco2022.femhub.com/account/abstracts/pdf/24/.

[32] Dzhoha A., Rozora I. Sequential resource allocation under multi-armed bandit model with online clustering as side information // Abstracts Baltic-Nordic-Ukrainian Workshop on Survey Statistics. — 2022. — Aug. — P. 42–43. — Access mode: https://wiki.helsinki.fi/display/BNU/Workshop+on+Survey+Statistics+2022+Scientific+Programme.

[33] Dzhoha A., Rozora I. The upper confidence bound strategy for multi-armed bandit problem // XXXVII International Conference «Problems of Decision Making under Uncertainties». — 2022. — P. 42–43. — Access mode: http://www.pdmu.univ.kiev.ua/PDMU2022/PDMU2022Sheki.pdf.

[34] Dzhoha A., Rozora I. Beta upper confidence bound policy for the design of clinical trials // Austrian Journal of Statistics. — 2023. — Aug. — Vol. 52, no. SI. — P. 26–39.

[35] Dzhoha A., Rozora I. Multi-armed bandit policy under delays for the design of clinical trial // Abstracts of 6th Baltic-Nordic-Ukrainian Conference on Survey Statistics. — 2023. — Aug. — P. 50–51. — Access mode: https://wiki.helsinki.fi/display/BNU/BANOCOSS2023.

[36] Dzhoha A., Rozora I. Multi-armed bandit problem with online clustering as side information // Journal of Computational and Applied Mathematics. — 2023. — Vol. 427. — P. 115132.

[37] Enabling Data-Driven Quality of Experience Optimization Using Group-Based Exploration-Exploitation / Jiang J., Sun S., Sekar V., and Zhang H. // 14th USENIX symposium on networked systems design and implementation (NSDI 17). — 2017. — P. 393–406.

[38] Faber V. Clustering and the continuous K-means algorithm // Los Alamos Science. — 1994. — Vol. 22.

[39] Feldman D. Contributions to the" two-armed bandit" problem // The Annals of Mathematical Statistics. — 1962. — Vol. 33, no. 3. — P. 847–856.

[40] Gambling in a rigged casino: The adversarial multi-armed bandit problem / Auer P., Cesa-Bianchi N., Freund Y., and Schapire R. E. // Proceedings of IEEE 36th annual foundations of computer science / IEEE. — 1995. — P. 322–331.

[41] Garivier A., Cappé O. The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond // Proceedings of the 24th Annual Conference on Learning Theory / ed. by Kakade S. M., von Luxburg U. — Budapest, Hungary : PMLR. — 2011. — Vol. 19 of Proceedings of Machine Learning Research. — P. 359–376. — Access mode: https://proceedings.mlr.press/v19/garivier11a.html.

[42] Gerchinovitz S., Lattimore T. Refined lower bounds for adversarial bandits // Advances in Neural Information Processing Systems. — 2016. — Vol. 29. — Access mode: https://proceedings.neurips.cc/paperfiles/paper/2016/file/2f37d10131f2a483a8dd005b3d14b0d9-Paper.pdf.

[43] Ghosh B. K., Sen P. K. Handbook of sequential analysis. — CRC Press, 1991.

[44] Gittins J. A dynamic allocation index for the sequential design of experiments // Progress in statistics. — 1974. — P. 241–266.

[45] Gittins J. C. Bandit processes and dynamic allocation indices // Journal of the Royal Statistical Society Series B: Statistical Methodology. — 1979. — Vol. 41, no. 2. — P. 148–164.

[46] Haldane J. On a method of estimating frequencies // Biometrika. — 1945. — Vol. 33, no. 3. — P. 222–225.

[47] Hartung J., Knapp G., Sinha B. K. Statistical meta-analysis with applications. — John Wiley & Sons, 2011.

[48] Hoeffding W. Probability inequalities for sums of bounded random variables // Journal of the American Statistical Association. — 1963. — Vol. 58, no. 301. — P. 13–30. — Access mode: http://www.jstor.org/stable/2282952 (online; accessed: 2023-05-29).

[49] Kaufmann E., Korda N., Munos R. Thompson Sampling: An Asymptotically Optimal Finite-Time Analysis // Algorithmic Learning Theory / ed. by Bshouty N. H., Stoltz G., Vayatis N., Zeugmann T. — Berlin, Heidelberg : Springer Berlin Heidelberg. — 2012. — P. 199–213.

[50] Koopman B. O. On distributions admitting a sufficient statistic // Transactions of the American Mathematical Society. — 1936. — Vol. 39. — P. 399–409. — Access mode: https://api.semanticscholar.org/CorpusID:122698932.

[51] Korda N., Kaufmann E., Munos R. Thompson sampling for 1-dimensional exponential family bandits // Advances in neural information processing

systems. — 2013. — Vol. 26.

[52] Kullback S., Leibler R. A. On information and sufficiency // The annals of mathematical statistics. — 1951. — Vol. 22, no. 1. — P. 79–86.

[53] Kullback–Leibler upper confidence bounds for optimal sequential allocation / Cappé O., Garivier A., Maillard O.-A., Munos R., and Stoltz G. // The Annals of Statistics. — 2013. — Vol. 41, no. 3. — P. 1516 – 1541.

[54] Lai T., Robbins H. Asymptotically efficient adaptive allocation rules // Advances in Applied Mathematics. — 1985. — Vol. 6, no. 1. — P. 4–22.

[55] Lai T. L. Adaptive Treatment Allocation and the Multi-Armed Bandit Problem // The Annals of Statistics. — 1987. — Vol. 15, no. 3. — P. 1091–1114. — Access mode: http://www.jstor.org/stable/2241818.

[56] Lattimore T. Optimally Confident UCB : Improved Regret for Finite-Armed Bandits // ArXiv. — 2015. — Vol. abs/1507.07880. — Access mode: https://api.semanticscholar.org/CorpusID:7771863.

[57] Li S., Karatzoglou A., Gentile C. Collaborative Filtering Bandits // Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. — Pisa Italy : ACM. — 2016. — P. 539–548.

[58] A near-optimal exploration-exploitation approach for assortment selection / Agrawal S., Avadhanula V., Goyal V., and Zeevi A. // Proceedings of the 2016 ACM Conference on Economics and Computation. — 2016. — P. 599–600.

[59] The nonstochastic multiarmed bandit problem / Auer P., Cesa-Bianchi N., Freund Y., and Schapire R. E. // SIAM journal on computing. — 2002. — Vol. 32, no. 1. — P. 48–77.

[60] O'neill M. E. PCG: A family of simple fast space-efficient statistically good algorithms for random number generation // ACM Transactions on Mathematical Software. — 2014.

[61] Online-Learning Congestion Control / Dong M., Meng T., Zarchy D., Arslan E., Gilad Y., Godfrey B., and Schapira M. // 15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18). — 2018. — P. 343–356.

[62] Pollard D. A User's Guide to Measure Theoretic Probability. Cambridge Series in Statistical and Probabilistic Mathematics. — Cambridge University Press, 2001. — P. 84–85.

[63] Puterman M. L. Markov Decision Processes: Discrete Stochastic Dynamic Programming. — 1st ed. — USA : John Wiley & Sons, Inc., 1994. — ISBN: 0471619779.

[64] R Core Team. — R: A Language and Environment for Statistical Computing. — R Foundation for Statistical Computing, Vienna, Austria, 2022. — Access mode: `https://www.R-project.org/`.

[65] Robbins H. Some Aspects of the Sequential Design of Experiments // Bulletin of the American Mathematical Society. — 1952. — Vol. 58, no. 5. — P. 527 – 535.

[66] Sauré D., Zeevi A. Optimal dynamic assortment planning with demand learning // Manufacturing & Service Operations Management. — 2013. — Vol. 15, no. 3. — P. 387–404.

[67] Siegmund D. Herbert Robbins and sequential analysis // The Annals of Statistics. — 2003. — Vol. 31, no. 2. — P. 349–365.

[68] Slivkins A. Introduction to Multi-Armed Bandits // Foundations and Trends® in Machine Learning. — 2019. — Vol. 12, no. 1-2. — P. 1–286.

[69] Snell J. L. Applications of martingale system theorems // Transactions of the American Mathematical Society. — 1952. — Vol. 73, no. 2. — P. 293–312.

[70] Stein C. A two-sample test for a linear hypothesis whose power is independent of the variance // The Annals of Mathematical Statistics. — 1945. — Vol. 16, no. 3. — P. 243–258.

[71] Stoltz G. Incomplete information and internal regret in prediction of individual sequences : Ph. D. thesis ; Université Paris Sud-Paris XI. — 2005.

[72] Thompson W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples // Biometrika. — 1933. — Vol. 25, no. 3-4. — P. 285–294.

[73] Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms / Li L., Chu W., Langford J., and Wang X. // Proceedings of the fourth ACM international conference on Web search and data mining. — 2011. — P. 297–306.

[74] Van Rossum G., Drake Jr F. L. Python Reference Manual. — Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[75] Varatharajah Y., Berry B. A Contextual-Bandit-Based Approach for Informed Decision-Making in Clinical Trials // Life. — 2022. — Vol. 12, no. 8.

[76] Villar S. S., Bowden J., Wason J. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges // Statistical science: a review journal of the Institute of Mathematical Statistics. — 2015. — Vol. 30, no. 2. — P. 199.

[77] Villar S. S., Rosenberger W. F. Covariate-adjusted response-adaptive randomization for multi-arm clinical trials using a modified forward looking Gittins index rule // Biometrics. — 2018. — Vol. 74, no. 1. — P. 49–57.

[78] Wald A. Sequential analysis. — Courier Corporation, 2004.

[79] Wald A., Wolfowitz J. Optimum character of the sequential probability ratio test // The Annals of Mathematical Statistics. — 1948. — P. 326–339.

[80] Whittle P. Restless bandits: Activity allocation in a changing world // Journal of applied probability. — 1988. — Vol. 25, no. A. — P. 287–298.

Додаток А

Список публікацій здобувача за темою дисертації та відомості про апробацію результатів дисертації

А.1. Список публікацій здобувача за темою дисертації

Публікації, в яких опубліковано основні наукові результати дисертації

1. Dzhoha A. S. Multi-armed bandit problem under delayed feedback // Bulletin of Taras Shevchenko National University of Kyiv. Series: Physics and Mathematics. 2021. no. 1, P. 20–26.
2. Dzhoha A. S. Sequential resource allocation in a stochastic environment: an overview and numerical experiments // Bulletin of Taras Shevchenko National University of Kyiv. Series: Physics and Mathematics. 2021. no. 3, P. 13–25.
3. Dzhoha A. S, Rozora I. V. Multi-armed bandit problem with online clustering as side information // Journal of Computational and Applied Mathematics. 2023. Vol. 427, P. 115–132.
4. Dzhoha A. S., Rozora I. V. Beta upper confidence bound policy for the design of clinical trials // Austrian Journal of Statistics. 2023. Vol. 52, no. SI, P. 26–39.

Публікації, які засвідчують апробацію матеріалів дисертації

5. Джога А. С. Модель багаторукого бандита у стохастичному середовищі та чисельні експерименти // VIII Всеукраїнська науково-практична конференція «Інформаційні технології — 2021. Математичне моделювання та обчислювальні методи». Київ, Україна. 20 травня 2021. С. 176–177.
6. Dzhoha A. S., Lebedev E. O. Sequential resource allocation under multi-armed bandit model with delays // XXXVI International Conference «Problems of Decision Making under Uncertainties». Kyiv, Ukraine. May 11-14, 2021. P. 37–38.
7. Dzhoha A. S., Rozora I. V. Multi-armed bandit problem with online clustering as side information // International Congress of Computational Engineering and Sciences ESCO. Pilsen, Czech Republic. June 13-16, 2022.
8. Dzhoha A. S., Rozora I. V. Sequential resource allocation under multi-armed bandit model with online clustering as side information // Baltic-Nordic-Ukrainian Workshop on Survey Statistics. Tartu, Estonia. August 23-26, 2022. P. 42–43.
9. Dzhoha A. S., Rozora I. V. The upper confidence bound strategy for multi-armed bandit problem // XXXVII International Conference «Problems of Decision Making under Uncertainties». November 23-25, 2022. P. 42–43.
10. Dzhoha A. S., Rozora I. V. Multi-armed bandit policy under delays for the design of clinical trial // 6th Baltic-Nordic-Ukrainian Conference on Survey Statistics. Helsinki, Finland. August 21-25, 2023. P. 50–51.

А.2. Відомості про апробацію результатів дисертації

Конференції

1. VIII Всеукраїнська науково-практична конференція «Інформаційні технології — 2021. Математичне моделювання та обчислювальні методи», 20 травня 2021 року.
2. «XXXVI International Conference Problems of Decision Making under Uncertainties», 11-14 травня 2021 року, секційна доповідь.
3. «8th International Congress of Computational Engineering and Sciences ESCO», Пльзень, Чеська Республіка, 13-17 червня 2022 року, секційна доповідь.
4. «Baltic-Nordic-Ukrainian Workshop on Survey Statistics», Тарту, Естонія, 23-26 серпня 2022 року, секційна доповідь.
5. «XXXVII International Conference Problems of Decision Making under Uncertainties», 23-25 листопада 2022 року, секційна доповідь.
6. «6th Baltic-Nordic-Ukrainian Conference on Survey Statistics», Гельсінкі, Фінляндія, 21-25 серпня 2023 року, секційна доповідь.

Додаток Б

Параметри середовищ та результати для жадібної стратегії

Таблиця Б.1

Параметри дій середовищ з бета-розподілом для експериментів з жадібною стратегією, значення сукупних втрат за результатами, агрегованими з 10000 незалежних тестів, та верхня границя згідно з теоремою 5.1 при використанні оптимізації (5.2)

$\Delta\mu$	Результати, $L(T)$	Верхня границя, $L(T)$	α_1	β_1	α_2	β_2
0.01	4.47	10.80	115	117	117	114
0.02	7.73	30.17	113	120	119	112
0.03	10.95	36.45	112	120	119	109
0.05	18.56	38.33	111	120	120	106
0.1	23.09	33.14	100	122	122	100
0.2	18.61	23.44	100	150	150	100
0.3	14.99	18.32	90	165	160	85
0.4	13.01	15.33	77	175	175	75
0.5	11.08	12.97	65	200	200	67
0.6	10.13	11.52	55	220	215	55
0.7	9.06	10.31	45	250	225	40
0.8	8.79	9.33	30	270	240	27