УДК 519.814, 519.816 https://doi.org/10.17721/1812-5409.2021/3.1

А. С. Джога, *аспірант*

Послідовний розподіл ресурсів у стохастичному середовищі: загальний опис, аналіз та чисельні експерименти

Київський національний університет імені Тараса Шевченка, 83000, м. Київ, пр-т. Глушкова 4д,
e-mail: andrew.djoga@gmail.com

A. S. Dzhoha, *PhD student*

Sequential resource allocation in a stochastic environment: an overview and numerical experiments

Taras Shevchenko National University of Kyiv, 83000, Kyiv, 4d Glushkova str.,
e-mail: andrew.djoga@gmail.com

У даній статті наводиться стислий огляд послідовного аналізу, а також місце в ньому послідовного розподілу ресурсів. Розглядаються стратегії послідовного розподілу ресурсів у середовищі, яке представлене моделлю стохастичного багаторукого бандита. Головною метою в представленій моделі є максимізація прибутку, що еквівалентно мінімізації втрат. У такому середовищі алгоритм послідовно вибирає дію та отримує винагороду з розподілу, пов'язаного з цією дією. Параметри такого розподілу не відомі заздалегідь. В статті наводиться огляд моделі багаторукого бандита у стохастичному середовищі та розглядаються стратегії для даного випадку. Наведені теоретичні результати використовуються для розробки програмного забезпечення для дослідження стратегій. Наведений чисельний експеримент.

Ключові слова: послідовний аналіз, проблема багаторукого бандита, мінімізація втрат.

In this paper, we consider policies for the sequential resource allocation under the multi-armed bandit problem in a stochastic environment. In this model, an agent sequentially selects an action from a given set and an environment reveals a reward in return. In the stochastic setting, each action is associated with a probability distribution with parameters that are not known in advance. The agent makes a decision based on the history of the chosen actions and obtained rewards. The objective is to maximize the total cumulative reward, which is equivalent to the loss minimization. We provide a brief overview of the sequential analysis and an appearance of the multi-armed bandit problem as a formulation in the scope of the sequential resource allocation theory. Multi-armed bandit classification is given with an analysis of the existing policies for the stochastic setting. Two different approaches are shown to tackle the multi-armed bandit problem. In the frequentist view, the confidence interval is used to express the exploration-exploitation trade-off. In the Bayesian approach, the parameter that needs to be estimated is treated as a random variable. Shown, how this model can be modelled with help of the Markov decision process. In the end, we provide numerical experiments in order to study the effectiveness of these policies.

Key Words: sequential analysis, multi-armed bandit problem, regret minimization.

Статтю представила д.ф.-м.н., доцент Розора І. В.

1 Вступ

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

Проблема багаторукого бандита представляє собою 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 і використанням є необхідним для досягнення оптимального результату з найменшими втратами в довгостроковій перспективі. Зрозуміло, що система, яка завжди вибирає дослідження нових варіантів, буде втрачати можливість отримання переваг від вже здобутих знань. З іншого боку, система, яка використовує тільки існуючі знання, не в змозі адаптуватися до значних змін у зовнішньому середовищі для досягнення оптимального ре-

зультату 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.

Сучасна історія послідовного аналізу починається з дослідження перевірки статистичних гіпотез, проведеного Wald A. [1], для знаходження методів отримання статистичних висновків з не фіксованим наперед числом випробувань контролю якості продукції. Теоретичні дослідження у послідовному аналізі та їх застосування знайшли відображення в роботах Barnard G. A., Anscombe F. J., Stein C., Wolfowitz J., Thompson W. R., Robbins H. та інших.

Успіх теорії послідовного аналізу в області перевірки статистичних гіпотез дав поштовх дослідженню послідовного оцінювання. Haldane J. B. S. [2] та Stein C. [3] описали, як деякі проблеми точкового та інтервального оцінювання можна розв'язувати за допомогою послідовного аналізу. У своїй роботі Stein C. продемонстрував, що таким чином можливо отримати довірчий інтервал з довжиною, яка задана експериментатором та не залежить від дисперсії розподілу.

Інша теорія, яка витікає з теорії послідовного аналізу і є фундаментальною для багатьох класів проблем багаторукого бандита, це теорія оптимальної зупинки. Каталізатором виокремлення цього напрямку послужили роботи Wald A. & Wolfowitz J. [4] та Arrow K. J., Blackwell D. & Girshick M. A. [5], присвяченні дослідженню задач послідовної перевірки статистичних гіпотез. В цих роботах був запропонований підхід Баєса для розв'язку проблеми оптимальної зупинки. 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 в очікуванні отримати більшу винагороду в майбутньому. В даному випадку оптимальне правило зупинки N можливо знайти через максимізацію очікуваної винагороди $\mathbb{E}[R_N]$. Для пошуку N можемо відштовхуватися від рівняння типу

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

Проблему оптимальної зупинки у загальному вигляді описав Snell J. L. [6], а Bellman R. [7] створив розділ математики, присвячений методам розв'язання багатокрокових задач оптимального керування — динамічне програмування.

Подальші дискусії, присвячені послідовному розподілу ресурсів, були започатковані у роботах Thompson W. R. [8] та Robbins H. [9], де автори аналізували проблему багаторукого бандита. Розглядався експеримент, де особа, що приймає рішення, на кожному кроці t послідовно обирає одну з двох дій (дворукий бандит), та спостерігає послідовність винагород $X_1, X_2, \ldots X_n$. З кожною дією пов'язаний розподіл імовірностей, параметри якого не відомі заздалегідь. Головною метою експерименту було пошук стратегії, при якій послідовний вибір дій приводив би до найбільшої можливої сукупної винагороди за n кроків $\sum_{t=1}^{n} X_t$. Результатом роботи стали опис математичної моделі проблеми багаторукого бандита та формалізація поняття ефективності цих стратегій. Ці дослідження послужили початком вивчення проблеми багаторукого бандита як окремого напрямку.

Проблему 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 відмовляються від усіх припущень щодо характеру процесів, пов'язаних з діями в моделі; та інші.

Також варто зазначити один з важливих проривів в цій області, зроблений у роботі Gittins J. & Jones D. M. [10]. Автори цієї роботи використовували теорію динамічного програмування з геометричним дисконтуванням для розв'язку задачі багаторукого бандита, яку описав Robbins H. Результатом роботи став узагальнений розв'язок проблеми багаторукого бандита за допомогою, так званих, індексів динамічного розподілу, що дозволило звести поставлену задачу до параметризованого сімейства розв'язків проблеми оптимальної зупинки.

Більш детальний огляд розвитку послідовного аналізу та опис ранніх робіт з проблеми багаторукого бандита, можна знайти у роботах Ghosh B. K. & Sen P. K. [11] та Siegmund D. [12].

За останні 10-15 років кількість досліджень

у цьому напрямку істотно зросла, що пов'язано з розвитком цифрових технологій та їх широкого використання. Наприклад, модель багаторукого бандита, як адаптивна стратегія, часто використовується у системах рекомендацій, динамічному ціноутворенні, адаптивній маршрутизації для мінімізації затримок у мережі та клінічних випробуваннях.

У даній роботі розглядається стратегія для послідовного розподілу ресурсів у середовищі, яке представлене моделлю стохастичного багаторукого бандита. У наступних розділах приводиться класифікація моделей багаторукого бандита, наводиться їх математична модель, дається аналіз існуючих стратегій та проводяться чисельні експерименти з метою вивчення ефективності цих стратегій.

2 Класифікація моделей багаторукого бандита

Проблема багаторуких бандитів — це послідовна взаємодія між суб'єктом, що приймає рішення, так званим агентом, та зовнішнім середовищем. Ця взаємодія відбувається протягом n кроків, де n — натуральне число, що називається горизонтом. На кожному кроці $t = 1, 2, \ldots, n$ агент обирає дію A_t із заданої множини $\mathcal{A} = \{1, 2, \ldots, k\}$, у відповідь середовище видає винагороду $X_t \in \mathbb{R}_{\geqslant 0}$. Вибір дії A_t залежить від історії попередніх виборів і їх результатів $H_{t-1} = (A_1, X_1, \ldots, A_{t-1}, X_{t-1})$. Метою агента є послідовний вибір таких дій із заданої множини $\mathcal{A}$, які призводять до найбільшої можливої сукупної винагороди за n кроків, тобто $\sum_{t=1}^{n} X_t$.

В даній постановці задачі стратегія агента — це відображення з множини виборів та їх результатів в множину дій, агент будує стратегію для 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.

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

Однією з метрик вимірювання ефективності стратегії агента є його втрати R_n за n кроків, що є різницею між очікуваною винагородою при виборі оптимальної дії на кожному кроці та винагородою при виборі дій відповідно до заданої стратегії. Вперше ця метрика була запропонована у роботі Lai T. L. & Robbins H. [13]. Сукупні втрати є функцією від часу та можуть бути визначені як втрати знецінювання на нескінченному горизонті, чи 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 n.

За різними властивостями проблему багаторукого бандита можна поділити на велику кількість категорій. За кількістю кроків розглядають моделі зі скінченним і нескінченним горизонтом. За кількістю дій проблему можна поділити на моделі з двома діями, k 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 (контексту).

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

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

Очікувані сукупні втрати при використанні стратегії s, яка визначає послідовність вибору дій $I_1^s, \ldots, I_n^s$ на горизонті n, визначається як

$$\mathbb{E}\left[R_n^s\right] = \mathbb{E}\left[\max_{i=1,\ldots,k} \sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{I_t^s,t}\right],$$

де $X_{i,t}$ це винагорода отримана від середовища на кроці t при виборі дії i. Ця модель буде розглядатися детально у наступних розділах даної статті.

Змагальний багаторукий бандит. 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 частіше використовуються втрати за найгіршим для обраної стратегії сценарієм. Нехай $\mathcal{P}$ — це множина усіх можливих послідовностей $k \times n$ на $[0,1]$ для усіх $n \in \mathbb{N}$, тоді очікувані втрати за найгіршим для обраної стратегії s сценарієм на послідовності $P \in \mathcal{P}$:

$$\sup_{P \in \mathcal{P}} \mathbb{E}\left[R_n^{s,P}\right].$$

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

Найменші очікувані втрати у найгіршому випадку, які може отримати будь-яка стратегія s з усіх можливих S, виражається через мінімакс втрати:

$$\inf_{s \in S} \sup_{P \in \mathcal{P}} \mathbb{E}\left[R_n^{s,P}\right].$$

Найпоширеніший алгоритм (імплементації однієї зі стратегій) для змагального багаторукого бандита є результатом роботи Auer P. та інших [14] — Exp3. Цей алгоритм використовує вагу у вигляді експоненціальної функції для оцінювання дій. Експонентне зростання допомагає значно збільшити вагу кращих дій.

Марковський багаторукий бандит. У марковських бандитах кожна дія асоційована з ланцюгом Маркова, а перехід до нового стану відбувається, коли цю дію вибирають. Ця модель була розглянута у роботі Gittins J. [15], 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 Whittle P. [16]. У даному випадку проблема не має загального розв'язку.

У стандартній марковській моделі дія $i = 1, 2, \ldots, k$ описується нерозкладним аперіодичним 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.

3 Модель стохастичного багаторукого бандита

Модель стохастичного багаторукого бандита задається k-вимірним вектором $\nu = (\nu_1, \ldots, \nu_k)$, де k – це кількість дій; кожна дія ν_i – це розподіл імовірностей з математичним сподіванням μ_i. На кожному кроці $t = 1, 2, ;n$ агент взаємодіє з моделлю, вибираючи дію A_t. У відповідь, модель виконує відбір вибірки $X_t \in \mathbb{R}_{\geqslant 0}$ з розподілу, пов'язаного з дією A_t та, як результат, реалізація вибірки стає доступною для агента. Вибірка $(X_1, \ldots, X_n)$ розглядається як винагорода, метою агента є максимізація сукупної винагороди за n кроків.

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

Одним з класів середовищ є модель багаторукого бандита з розподілом Бернуллі. Модель цьому класу описується вектором середніх значень $\mu \in [0,1]^k$, де кожен елемент μ_i – це параметр розподілу Бернуллі дії i. Для моделі цього класу X_t є випадковою величиною, пов'язаною з розподілом імовірностей μ_{A_t}, тобто $X_t \sim \text{Bernoulli}(\mu_{A_t})$ є винагородою агента на кроці t при виборі дії A_t.

Якщо розглядається параметрична модель, то розподіл ν_i, відповідно, залежить від деякого невідомого параметра θ_i, який набуває значення з множини Θ. Нехай $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \in \Theta^k$, тоді параметрична модель стохастичного багаторукого бандита задається як $\nu = \nu_{\boldsymbol{\theta}} = (\nu_{\theta_1}, \ldots, \nu_{\theta_k})$. Позначимо через μ^* середнє значення розподілу ймовірностей оптимальної дії, тобто $\mu^* = \max_{i \in \mathcal{A}}(\mu_i)$. Тоді, згідно з [13], для параметричної моделі стохастичного багаторукого бандита можемо описати очікува-

ні сукупні втрати:

$$R_n^s(\theta) = \mathbb{E}_\theta\left[n\mu^* - \sum_{t=1}^{n} X_t\right] = n\mu^* - \mathbb{E}_\theta\left[\sum_{t=1}^{n} X_t\right], \tag{3.1}$$

які визначені для будь-якої стратегії s на горизонті n з параметром θ, який є фіксованим і невідомим. Головною метою даної моделі є максимізація винагороди, що еквівалентно мінімізації втрат.

У своїй роботі Lai T. L. & Robbins H. через введення фільтрації представили декомпозицію втрат, як функцію від кількості разів $N_i(n)$ кожна дія i була вибрана на горизонті n, використовуючи стратегію s.

Введемо фільтрацію

$$\mathcal{F}_t = \sigma\left(A_1, X_1, \ldots, A_t, X_t\right),$$

де для недетермінованої стратегії відбір A_t здійснюється з розподілу ймовірностей p_t на системі дій $\{1, \ldots, k\}$, який є $\mathcal{F}_{t-1}$-вимірним. Використовуючи $\mathbb{E}\left[X_t|\mathcal{F}_{t-1}\right] = \mu_{A_t}$, отримаємо декомпозицію втрат у вигляді очікуваної суми неоптимальностей вибраних дій при використанні стратегії s:

$$R_n^s(\theta) = \mathbb{E}_\theta\left[\sum_{t=1}^{n}\left(\mu^* - \mu_{A_t}\right)\right] = \sum_{i\in\mathcal{A}}\left(\mu^* - \mu_i\right)\mathbb{E}_\theta\left[N_i(n)\right], \tag{3.2}$$

де $N_i(n) = \sum_{t=1}^{n} \mathbf{1}_{(A_t=i)}$.

За визначенням втрат R_n^s можна виділити їх наступні властивості, які є дійсними для будь-якої стратегії s:

1) $R_n^s \geqslant 0$, що випливає з декомпозиції втрат (3.2), де $\mu^* \geqslant \mu_i$ та $N_i(n) \geqslant 0$ за визначенням.

2) Якщо виконується $R_n^s = 0$, це свідчить, що агент знає оптимальну дію заздалегідь, тобто $\mathbb{P}\left(\mu_{A_t} = \mu^*\right) = 1$ для всіх $t = 1, 2, \ldots n$.

3) З формули (3.1) видно, що $R_n^s \leqslant n\mu^*$.

Якщо ми припустимо, що винагорода приймає значення в замкненому проміжку $X_t \in [0, 1]$, тоді маємо $R_n^s \leqslant n$. Тоді для будь-якої стратегії s маємо $R_n^s = O(n)$. Таким чином, наша головна мета це знаходження стратегій, для яких втрати як мінімум сублінійні, тобто $R_n^s = o(n)$.

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

$$x' = \frac{x - a}{b - a}.$$

4 Асимптотичний аналіз втрат

Розглянемо параметричні моделі з розподілами з експоненційного сімейства [17], тобто розподіл дії i залежить від параметра θ_i. За допомогою функцій $\eta(\theta)$, $T(x)$, $A(\theta)$, $h(x)$ щільність може бути подана у вигляді

$$f_X(x; \theta_i) = exp\left(\eta(\theta_i) \cdot T(x) + A(\theta_i)\right) h(x).$$

Це дозволяє використовувати розходження Кульбака-Лейблера [18] (також називають відносною ентропією) як міру того, наскільки розподіл p однієї дії відрізняється від іншої q:

$$\mathrm{KL}(p, q) = \int \log\left(\frac{\mathrm{d}p}{\mathrm{d}q}(x)\right) \mathrm{d}x.$$

Для розподілу Бернуллі, який входить в експоненційне сімейство, розходження Кульбака-Лейблера має вигляд

$$\mathrm{KL}(p, q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q},$$

де p, q — це параметри розподілів відповідно.

Представлені у роботах Lai T. L. & Robbins H. у роботах [9, 13] перші стратегії характеризувались фіксованою кількістю досліджень простору варіантів та використанням найоптимальнішого варіанту за вибірковим середнім. Автори показали, що для ціх стратегій можливо отримати сублінійні втрати:

$$\lim_{n\to\infty}\frac{R_n^s}{n} = 0.$$

За результатами були сформульовані наступні твердження:

Означення 4.1. Стратегія s є рівномірно ефективною, якщо її втрати задовольняють $R_n^s(\theta) = o(n^\alpha)$ для всіх $\alpha \in (0, 1]$.

Для параметричних моделей зі скалярним параметром, тобто $\Theta \in R$, було показано, що для рівномірно ефективних стратегій кількість виборів кожної неоптимальної дії i ($\mu_i < \mu^*$) має як мінімум логарифмічну складність:

$$\liminf_{n\to\infty} \frac{\mathbb{E}_\theta[N_i(n)]}{\log(n)} \geqslant \frac{1}{\mathrm{KL}(\nu_{\theta_i}, \nu_{\theta^*})}.$$

Для моделі з розподілом Бернуллі за допомогою декомпозиції втрат (3.2) для рівномірно ефективної стратегії s можна отримати наступну нерівність [19]:

$$\liminf_{n\to\infty} \frac{R_n^s}{\log(n)} \geqslant \sum_{i:\Delta_i>0} \frac{\Delta_i}{\mathrm{KL}(\mu_i, \mu^*)},$$

де $\Delta_i = \mu^* - \mu_i$.

Також у [20] для моделі з розподілом Бернуллі було показано, що жодна зі стратегій не може отримати втрати $R_n^s = o(\log(n))$.

Таким чином маємо, що для будь-якої стратегії у будь-якому класі стохастичної моделі втрати становлять що найменше $\Omega(\log(n))$.

Означення 4.2. Якщо для стратегії s нижня і верхня границі співпадають згідно до асимптотичної поведінки функцій (нотація Ландау), то її називають асимптотично оптимальною.

Тобто для рівномірно ефективної стратегії s для моделі з розподілом Бернуллі, стратегія вважається асимптотично оптимальною (4.2), якщо виконується нерівність

$$\limsup_{n\to\infty} \frac{R_n^s}{\log(n)} \leqslant \sum_{i:\Delta_i>0} \frac{\Delta_i}{\mathrm{KL}(\mu_i, \mu^*)},$$

де $\Delta_i = \mu^* - \mu_i$.

У стохастичній моделі зазвичай розглядається асимптотичний аналіз втрат у визначеному класі середовища, тобто нижня границя втрати залежить від розподілу дій. Також можливо визначити критерій мінімаксу як у змагальних бандитах, тобто показати нижню границю втрат без залежності від розподілу дій. Bubeck S. & Cesa-Bianchi N. [19] показали, що для будь-якої стратегії s існує стохастична модель багаторукого бандита ν, де виконується

$$R_n^s(\nu) \geqslant \frac{1}{20}\sqrt{kn},$$

де k — це кількість дій.

5 Стратегії для стохастичного багаторукого бандита

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. Наприклад, стратегія з рівномірним дослідженням дій буде ефективною, коли усі дії з заданої множини $\mathcal{A} = \{1, 2, \ldots, k\}$ є оптимальними, тобто $\mu_1 = \mu_2 = \cdots = \mu_k$. Інакше втрати набувають лінійну складність, тобто $R_n^s = O(n)$. Це можна показати через декомпозицію втрат (3.2):

$$R_n = \sum_{i=1}^{k} (\mu^* - \mu_i)\, \mathbb{E}\left[N_i(n)\right],$$

де $N_i(n)$ — це кількість разів дія i була вибрана за n кроків. Тоді

$$\mathbb{E}\left[N_i(n)\right] = \mathbb{E}\left[\sum_{t=1}^{n} \mathbf{1}_{(A_t=i)}\right] = n\,\mathbb{P}\left(A_t = i\right) = \frac{n}{k}.$$

Отже, для стохастичного багаторукого бандита з кількістю дій $k > 1$ та принаймні однією неоптимальною дією, стратегія з рівномірним дослідженням має втрати

$$R_n = \frac{n}{k} \sum_{i=1}^{k} \Delta_i,$$

де $\Delta_i = \mu^* - \mu_i$ та $\mu^* \geqslant \mu^i$.

З іншого боку, стратегія, яка використовує тільки існуючі знання, нехтуючи дослідженням можливих змін у середовищі може понести значні втрати. Наприклад, стратегія моделі стохастичного дворукого бандита ($k = 2$), яка за перші два кроки обирає першу та другу дії відповідно, а на всіх інших $t \in \{3, 4, \ldots, n\}$ використовує найкращу дію за вибірковим середнім, тобто $\arg\max_{i\in\{1,2\}} \hat{\mu}_i(t)$, де

$$\hat{\mu}_i(t) = \frac{1}{N_i(t)} \sum_{j=1}^{t} \mathbf{1}_{(A_j=i)} X_j,$$

може отримати лінійні втрати. Для зразку візьмемо розподіли Бернуллі з параметрами

$p < 1/2$ та $q = 1/2$ для першої та другої дій відповідно. Тоді, з імовірністю pq, стратегія отримає винагороди 1 та 0 за першу та другу дії відповідно, і буде змушена обирати неоптимальну першу дію на всіх інших кроках. Звідси, згідно з визначенням (3.2), втрати набувають вигляду

$$R_n \geqslant \left(\frac{1}{2} - p\right)(n-1).$$

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

Стратегія Explore-First. Перші приклади стратегії Explore-First з'явилися у роботі Robbins H. [9], де він показав, що за деяких умов втрати можуть бути сублінійні.

Ця стратегія характеризується фіксованою кількістю m досліджень кожної дії в першій фазі дослідження, і вибором емпірично кращої дії у другій фазі для використання. Алгоритм вибору дії A_t на кроці t:

$$A_t = \begin{cases} (t \bmod k) + 1, & t \leqslant mk \\ \arg\max_{i \in \mathcal{A}} \hat{\mu}_i(mk), & t > mk. \end{cases}$$

Алгоритм стратегії Explore-First виглядає наступним чином:

1) фаза дослідження простору варіантів (перші mk кроків): вибрати кожну дію $i \in \mathcal{A}$ m разів;

2) фаза вибору оптимального варіанту ($t \in \{mk+1, \ldots, n\}$): використання $\arg\max_{i \in \mathcal{A}} \hat{\mu}_i(mk)$.

Якщо горизонт n відомий наперед, ми можемо мінімізувати верхню границю для отримання сублінійних втрат. Для моделі з розподілом Бернуллі Slivkins. A. [20] показав, що границя зверху є

$$\mathbb{E}[R_n] \leqslant n^{2/3} \times O\left(k \log n\right)^{1/3},$$

при виборі m наступним чином:

$$m = (n/k)^{2/3} \cdot O\left(\log n\right)^{1/3}. \qquad (5.1)$$

У [21] була отримана верхня границя з залежністю від неоптимальності дій:

$$R_n \leqslant m \sum_{i=1}^{k} \Delta_i + (n - mk) \sum_{i=1}^{k} \Delta_i \exp\left(-m\Delta_i^2\right).$$

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

Варіацією цієї стратегії без залежності від кількості кроків n є Epsilon-Greedy, яка займається дослідженням та використанням протягом усього горизонту.

Стратегія Epsilon-Greedy. Ця стратегія задається за допомогою деякого параметра $0 < \varepsilon < 1$, де на кожному кроці з імовірністю ε відбувається дослідження варіантів, вибираючи рівномірно випадково, та з імовірністю $1 - \varepsilon$ використовується найкраща дія за вибірковим середнім:

$$A_t = \begin{cases} i \sim \text{Uniform}(\{1, k\}), & \text{з імовірністю } \varepsilon \\ \arg\max_{i \in \mathcal{A}} \left(\hat{\mu}_i(t)\right), & \text{інакше.} \end{cases}$$

Втрати цієї стратегії лінійні, так як ми змушенні продовжувати досліджувати варіанти, тобто границя знизу становить що найменше:

$$R_n \geqslant \left(\varepsilon \frac{1}{k} \sum_{i \in \mathcal{A}} \Delta_i\right) n.$$

У [22] було показано, що можливо отримати втрати $O(k \log(n))$, якщо замість сталого значення ε використовувати спадну послідовність $(\varepsilon_t)_{t \in \mathbb{N}}$, але для цього потрібно знати наперед значення $\min_{i \in \mathcal{A}} \Delta_i$ для вибору кроку послідовності.

Стратегія Upper Confidence Bound (UCB). Ця стратегія з використанням принципу «оптимізму в умовах невизначеності»була представлена у роботі Auer P. та інші [22].

Для всіх дій $i \in \mathcal{A}$ обчислюється значення $U_i(t)$ на базі вибіркового середнього та верхньої границі надійного інтервалу, яка з великою ймовірністю є завищеною оцінкою невідомого математичного сподівання розподілу, пов'язаного з дією i.

На початку алгоритму кожна дія вибирається один раз, а потім на кожному кроці t — дія з найбільшим значенням $U_i(t)$:

$$A_t = \begin{cases} (t \bmod k) + 1, & t \leqslant k \\ \arg\max_{i \in \mathcal{A}} U_i(t), & t > k. \end{cases}$$

Обчислення верхніх границь будується на використанні нерівності Хефдинга [23]:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{t=1}^{n}\left(\mathbb{E}\left[X_t\right]-X_t\right)\right|\geqslant\xi\right)\leqslant\exp\left(-2n\xi^2\right),$$

де $X_1,\ldots,X_n$ це незалежно однаково розподілені випадкові величини з $X_t\in[0,1]$ для усіх $\xi\geqslant 0$. Ця нерівність може використовуватися для аналізу надійного інтервалу, де ліва частина відображає ймовірність похибки для надійного інтервалу навколо математичного сподівання з радіусом ξ.

Таким чином, для стохастичного багаторукого бандита маємо наступну нерівність:

$$\mathbb{P}\left(\mu_i\geqslant\hat{\mu}_i+\xi_i(t)\right)\leqslant\exp\left(-2n\xi_i(t)^2\right),$$

де для $\exp\left(-2n\xi_i(t)^2\right)$ бажано отримати найменше значення. Приймемо $\exp\left(-2n\xi_i(t)^2\right)=p$, тоді радіус надійного інтервалу (чи межа похибки) приймає вигляд:

$$\xi_i(t)=\sqrt{\frac{-\log(p)}{2N_i(t)}}.$$

Для стратегії UCB значення $U_i(t)=U_i^{\mathrm{UCB}}(t)$ обчислюється як сума вибіркового середнього $\hat{\mu}_i(t)$ та радіусу надійного інтервалу $\xi_i(t)$ для деякого параметра α:

$$U_i^{\mathrm{UCB}}(t)=\frac{1}{N_i(t)}\sum_{j=1}^{t}\mathbf{1}_{(A_j=i)}X_j+\sqrt{\alpha\frac{\log(t)}{N_i(t)}}.$$

В цьому рівнянні перша частина (вибіркове середнє $\hat{\mu}_i(t)$) відповідає за вибір кращої дії на даний час, а межа похибки $\xi_i(t)$ — за дослідження. Вибір дії i відбувається, якщо $\hat{\mu}_i(t)$ достатньо велике, що може свідчити про можливу оптимальність дії, та/або, якщо межа похибки завелика, це може означати, що дія i недостатньо досліджена. Параметр α допомагає контролювати цей баланс.

Для параметричних моделей з розподілами з експоненційного сімейства альтернативою використанню радіусу надійного інтервалу з нерівності Хефдинга є використання нерівності Чернова [24], як було зроблено у роботах [25, 26], де стратегія отримала назву KL-UCB (*англ.* Kullback–Leibler divergence Upper Confidence Bound).

За стратегією KL-UCB для всіх дій обчислюється значення $U_i(t)=U_i^{\text{KL-UCB}}(t)$ наступним чином:

$$U_i^{\text{KL-UCB}}(t)=\sup_{q\in[0,1]}\left\{q:\mathrm{KL}\left(\hat{\mu}_i,q\right)\leqslant\frac{\log(t)}{N_i(t)}\right\}.$$

Як видно з визначення $U_i^{\text{KL-UCB}}(t)$, для реалізації алгоритму потрібно розв'язати задачу оптимізації.

Втрати цих стратегій мають логарифмічний порядок, вони є асимптотично оптимальними.

6 Баєсова модель

З точки зору баєсової інтерпретації ймовірності можна розглядати параметр $\boldsymbol{\theta}$ як випадкову величину з відбору вибірки з деякого апріорного розподілу Π. Стратегія s, яка максимізує винагороди, також еквівалентно мінімізує баєсові втрати, які були представлені у [13]:

$$\begin{aligned}\mathrm{BR}_n^s\left(\Pi\right)&=\mathbb{E}_\Pi\left[n\mu^*-\sum_{t=1}^{n}X_t\right]\\&=\mathbb{E}_\Pi\left[\mathbb{E}_\Pi\left[n\mu^*-\sum_{t=1}^{n}X_t|\boldsymbol{\theta}\right]\right]\\&=\mathbb{E}_\Pi\left[R_n^s\left(\boldsymbol{\theta}\right)\right].\end{aligned}$$

На кроці t вибір дії заснований на поточному апостеріорному розподілу параметра $\boldsymbol{\theta}$, який є умовним розподілом, що залежить від поточних спостережень, та задається як

$$\Pi_t(\boldsymbol{\theta})=L(\boldsymbol{\theta}|A_1,X_1,\ldots,A_t,X_t),$$

де L — це функція вірогідності.

Bubeck S. & Cesa-Bianchi N. [19] показали, що для кожної стратегії s, існує такий апріорний розподіл Π, що дозволяє отримати наступну складність втрат:

$$\mathrm{BR}_n^s\left(\Pi\right)\geqslant\frac{1}{20}\sqrt{kn},$$

де k — це кількість дій.

Одна з перших баєсових стратегій для стохастичного багаторукого бандита — це Thompson Sampling [8]. Стратегія передбачає використання деякого k-вимірного вектора апріорного розподілу $\Pi_0=\left(\pi_0^1,\ldots,\pi_0^k\right)$, де π_0^i є апріорним розподілом параметра θ_i для дії i. На кроці t, враховуючи нові спостереження, виводиться новий апостеріорний розподіл Π_t:

$$\Pi_t=\left(\pi_t^1,\ldots,\pi_t^k\right),$$

де π_t^i це апостеріорний розподіл параметра θ_i для дії i після $m = N_i(t)$ спостережень $Y_{i,1}, \ldots, Y_{i,m}$:

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

Відбувається відбір вибірки $Z_{i,t} \sim \pi_t^i$ для кожної дії та аналізуються їх реалізації $(z_{1,t}, \ldots, z_{k,t})$ для вибору дії з найбільшим значенням:

$$A_t = \arg\max_{i \in \mathcal{A}} z_{i,t}.$$

Для стохастичного багаторукого бандита з розподілом Бернуллі в якості апріорного розподілу можна обрати Бета-розподіл, який задається параметрами $\alpha > 0$ та $\beta > 0$ для кожної дії i у наступному вигляді:

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

Таким чином, маємо на початку алгоритму рівномірний розподіл Uniform$(0, 1)$ для вибору дій випадковим чином. На кроці t виводиться наступний апостеріорний розподіл:

$$\pi_t^i = \text{Beta}(T_i(t) + 1, N_i(t) - T_i(t) + 1),$$

де $T_i(t)$ це сума усіх винагород дії i після t кроків:

$$T_i(t) = \sum_{j=1}^{t} \mathbf{1}_{(A_j = i)} X_j.$$

Значення $T_i(t)$ показує кількість успіхів від вибору дії i, а $(T_i(t) - N_i(t))$ — кількість невдач протягом горизонту t.

Agrawal S. & Navin G. [27] показали, що втрати стратегії Thompson Sampling для стохастичного багаторукого бандита у загальному випадку є логарифмічними:

$$R_n = O\left(\left[\left(\sum_{i=1}^{k} \frac{1}{{\Delta_i}^2}\right)^2\right] \log(n)\right).$$

Цей алгоритм є асимптотично оптимальним.

Також варто зазначити, що проблему багаторукого бандита у баєсовій інтерпретації можливо розв'язувати за допомогою динамічного програмування. Bellman R. [28] використовую для цього втрати знецінювання. Для деякого коефіцієнта знецінювання $\gamma \in (0, 1]$ потрібно максимізувати наступні сукупні винагороди:

$$\mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} X_t\right].$$

Було показано, що розв'язок за допомогою динамічного програмування [15, 29] є оптимальною стратегією у даному випадку. Також у баєсовій інтерпретації оптимальним є використання індексів динамічного розподілу (індексів Гіттінса), в основі яких лежить динамічне програмування, але це передбачає розглядання нескінченного горизонту. У моделі зі скінченним горизонтом індекси Гіттінса не є оптимальними.

Опис та аналіз стратегій для інших розподілів та їх варіацій можна знайти у таких роботах, як Bubeck S. & Cesa-Bianchi N. [19] та Slivkins. A. [20].

7 Чисельні експерименти

У цьому розділі представлені результати емпіричних тестів для стохастичного багаторукого бандита з розподілом Бернуллі. Для цього було розроблено програмне забезпечення з імплементацією алгоритмів для усіх розглянутих стратегій [33]. Мета експериментів — показати переваги наведених стратегій.

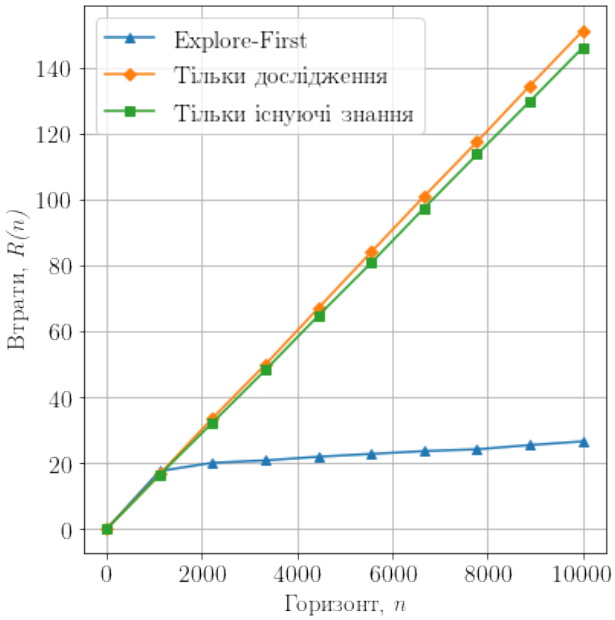

Рис. 1: Втрати багаторукого бандита Бернуллі зі стратегіями Explore-First (параметр m обран за (5.1)), з використанням тільки існуючого знання та тільки дослідження (рівномірного)

Експеримент проведено з горизонтом $n = 10000$ з двома діями і математичними очікуваннями 0.7 і 0.8 відповідно, результати агреговані на 1000 незалежних тестів. На рисунках 1 та 2 зображені графіки втрат розглянутих стратегій з порівнянням випадків вибору тільки дослі-

дження (рівномірного) та тільки використання існуючого знання (Explore-First з параметром $m = 1$)).

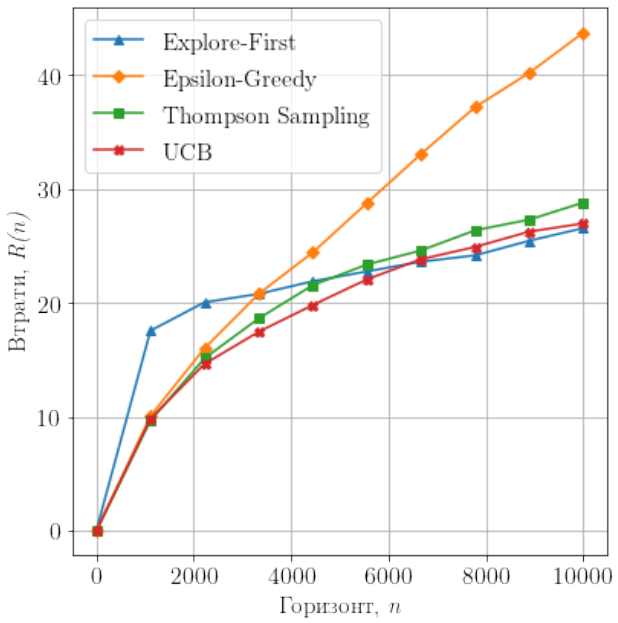

Рис. 2: Втрати багаторукого бандита Бернуллі зі стратегіями Explore-First (параметр m обран за (5.1)), Epsilon-Greedy з параметром $\varepsilon = 0.2$, Thompson Sampling та UCB з параметром $\alpha = 0.1$

8 Зв'язок з марковським процесом прийняття рішень

У загальному випадку модель багаторукого бандита може бути описана за допомогою марковського процесу прийняття рішень (МППР) [30].

Позначимо через $\Delta(G)$ множину всіх імовірнісних розподілів групи G, через S та $\mathcal{A}$ — 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 $i \in \mathcal{A}$ 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 $(S, \mathcal{A}, P, R)$.

Коли агент знаходиться у стані $s \in S$ та вибирає дію $i \in \mathcal{A}$, 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 3.

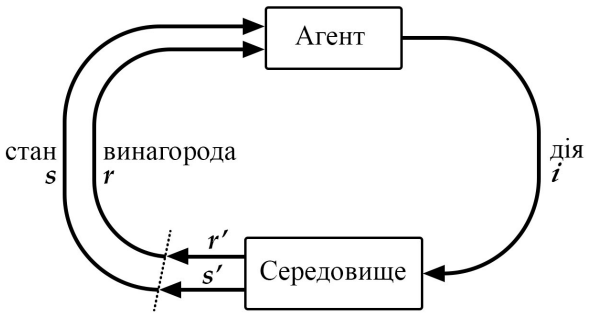

Рис. 3: Взаємодія агента зі стохастичним середовищем

Стохастична модель багаторукого бандита може бути описана як МППР з одним станом s_0, тоді для стохастичного ядра, яке задає винагороди, маємо $R(\cdot|s_0, i) = \nu_i$.

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

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

$$V_\Pi = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} X_t\right].$$

Знаходження оптимальної політики у МППР може вимагати забагато обчислення у зв'язку з потенційно великим простором станів. Розв'язки для випадку дворукого бандита були наведені у роботах [31, 32].

9 Висновки

Були розглянуті стратегії послідовного розподілу ресурсів у середовищі, яке представлене моделлю стохастичного багаторукого бандита з наведенням огляду та класифікації. Головною метою розглянутих стратегій є максимізація прибутку, що еквівалентно мінімізації

втрат. Проведено аналіз складності цих стратегій, умов їх ефективності та втрат для випадку багаторукого бандита у стохастичному середовищі з розподілом Бернуллі. Розглянуті теоретичні відомості використовувалися для розробки програмного забезпечення для проведення чисельних експериментів з метою демонстрації ефективності стратегій.

Список використаних джерел

1. *Wald A.* Sequential Analysis / A. Wald. — NY: John Wiley & Sons, Inc., 1950.
2. *Haldane J. B. S.* On a method of estimating frequencies / J. B. S. Haldane // Biometrika. — 1945. — Vol. 33. — No. 3. — P. 222–225.
3. *Stein C.* A two-sample test for a linear hypothesis whose power is independent of the variance / C. Stein // The Annals of Mathematical Statistics. — 1945. — Vol. 16. — No. 3. — P. 243–258.
4. *Wald A.* Optimum character of the sequential probability ratio test / A. Wald, J. Wolfowitz // Ann. Math. Statist. — 1948. — P. 326–339.
5. *Arrow K. J.* Bayes and minimax solutions of sequential decision problems / K. J. Arrow, D. Blackwell, M. A. Girshick // Econometrica. — 1949. — Vol. 17. — P. 213–244.
6. *Snell J. L.* Applications of martingale system theorems / J. L. Snell // Trans. Amer. Math. Soc. — 1952. — Vol. 73. — P. 293–312.
7. *Bellman R.* Dynamic Programming / R. Bellman. — Princeton Univ. Press, 1957.
8. *Thompson W. R.* On the likelihood that one unknown probability exceeds another in view of the evidence of two samples / W. R. Thompson // Biometrika. — 1933. — Vol. 25. — No. 3/4. — P. 285–294.
9. *Robbins H.* Some aspects of the sequential design of experiments / H. Robbins // Bulletin of the American Mathematical Society. — 1952. — Vol. 58. — No. 5. — P. 527–535.
10. *Gittins J.* A dynamic allocation index for the sequential design of experiments / J. Gittins, D. M. Jones // Progress in Statistics. — 1974. — P. 241–266.
11. *Ghosh B. K.* Handbook of sequential analysis / B. K. Ghosh. — CRC Press, 1991.
12. *Siegmund D.* Herbert Robbins and sequential analysis / D. Siegmund // Annals of statistics. — 2003. — P. 349–365.
13. *Lai T. L.* Asymptotically efficient adaptive allocation rules / T. L. Lai, H. Robbins // Advances in applied mathematics. — 1985. — Vol. 6. — No. 1. — P. 4–22.
14. *Auer P.* The nonstochastic multiarmed bandit problem / P. Auer, N. Cesa-Bianchi, Y. Freund, R. Schapire // SIAM Journal on Computing. — 2003. — Vol. 32. — No. 1. — P. 48–77.
15. *Gittins J.* Bandit processes and dynamic allocation indices / J. Gittins // Journal of the Royal Statistical Society. — 1979. — Vol. 41. — No. 2. — P. 148–177.
16. *Whittle P.* Restless bandits: Activity allocation in a changing world / P. Whittle // Journal of Applied Probability. — 1988. — Vol. 25. — P. 287–298.
17. *Koopman B. O.* On distributions admitting a sufficient statistic / B. O. Koopman // Transactions of the American Mathematical society. — 1936. — Vol. 39. — No. 3. — P. 399–409.
18. *Kullback S.* On information and sufficiency / S. Kullback, R. A. Leibler // The annals of mathematical statistics. — 1951. — Vol. 22. — No. 1. — P. 79–86.
19. *Bubeck S.* Regret analysis of stochastic and nonstochastic multi-armed bandit problems / S. Bubeck, N. Cesa-Bianchi // Foundations and Trends in Machine Learning. — 2012. — Vol. 5 — No. 1. — P. 1–122.
20. *Slivkins A.* Introduction to multi-armed bandits / A. Slivkins // Foundations and Trends in Machine Learning. — 2019. — Vol. 12. — No. 1-2. — P. 1–286.
21. *Джога А.* Багаторукий бандит з розподілом Бернуллі в середовищі з затримками /

А. Джога // Вісник Київського національного університету імені Тараса Шевченка. Серія: фізико-математичні науки. — 2021. — №. 1. — С. 20–26.

22. *Auer P.* Finite-time Analysis of the Multiarmed Bandit Problem / P. Auer, N. Cesa-Bianchi, P. Fischer // Machine Learning. — 2002. — Vol. 47. — No. 2. — P. 235–256.

23. *Hoeffding W.* Probability inequalities for sums of bounded random variables / W. Hoeffding // Journal of the American statistical association. — 1963. — Vol. 58. — No. 301. — P. 13–30.

24. *Chernoff H.* A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations / H. Chernoff // The Annals of Mathematical Statistics. — 1952. — Vol. 23. — No. 4. — P. 493–507.

25. *Garivier A.* The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond / A. Garivier, O. Cappe // Conference on Learning Theory, PMLR. — 2011. — P. 359–376.

26. *Cappe O.* Kullback-Leibler Upper Confidence Bounds For Optimal Sequential Allocation / O. Cappe, A. Garivier, O. A. Maillard, R. Munos, and G. Stoltz // Annals of Statistics. — 2013. — Vol. 41. — No. 3. — P. 1516–1541.

27. *Agrawal S.* Analysis of thompson sampling for the multi-armed bandit problem / S. Agrawal, N. Goyal // Conference on learning theory, JMLR. — 2012. — P. 39.

28. *Bellman R.* A problem in the sequential design of experiments / R. Bellman // The indian journal of statistics. — 1956. — Vol. 16. — No. 3/4. — P. 221–229.

29. *Berry D.* Bandit Problems. Sequential allocation of experiments / D. Berry, B. Fristedt // Chapman and Hall, Springer. — 1985. — Vol. 5. — No. 51-87. — P. 7.

30. *Puterman M.* Markov Decision Processes: discrete stochastic dynamic programming / M. Puterman. — Wiley, 1994.

31. *Feldman D.* Contributions to the "two-armed bandit"/ D. Feldman // The Annals of Mathematical Statistics. — 1962. — Vol. 33. — No. 3. — P. 947–956.

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

33. *Dzhoha A.* Multi-armed bandit problem under delayed feedback: numerical experiments [Електронний ресурс] / A. Dzhoha. — 2021. — Режим доступу до ресурсу: https://github.com/djo/delayed-bandit.

References

1. WALD, A. (1950) *Sequential Analysis*. John Wiley & Sons, Inc., NY.

2. HALDANE, J. B. S. (1945) On a method of estimating frequencies. *Biometrika*. 33 (3). p. 222–225.

3. STEIN, C. (1945) A two-sample test for a linear hypothesis whose power is independent of the variance. *The Annals of Mathematical Statistics*. 16 (3). p. 243–258.

4. WALD, A., WOLFOWITZ, J. (1948) Optimum character of the sequential probability ratio test. *Ann. Math. Statist*. p. 326–339.

5. ARROW, K. J., BLACKWELL, D., GIRSHICK, M. A. (1949) Bayes and minimax solutions of sequential decision problems. *Econometrica*. 17. p. 213–244.

6. SNELL, J. L. (1952) Applications of martingale system theorems. *Trans. Amer. Math. Soc*. 73. p. 293–312.

7. BELLMAN, R. (1957) *Dynamic Programming*. Princeton Univ. Press.

8. THOMPSON, W. R. (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*. 25 (3/4). p. 285–294.

9. ROBBINS, H. (1952) Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*. 58 (5). p. 527–535.

10. GITTINS, J., JONES, D. M. (1974) A dynamic allocation index for the sequential design of experiments. *Progress in Statistics*. p. 241–266.

11. GHOSH, B. K. (1991) *Handbook of sequential analysis.* CRC Press.

12. SIEGMUND, D. (2003) Herbert Robbins and sequential analysis. *Annals of statistics.* p. 349–365.

13. LAI, T. L., ROBBINS, H. (1985) Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics.* 6 (1). p. 4–22.

14. AUER, P., CESA-BIANCHI, N., FREUND, Y., SCHAPIRE, R. (2003) The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing.* 32 (1). p. 48–77.

15. GITTINS, J. (1979) Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society.* 41 (2). p. 148–177.

16. WHITTLE, P. (1988) Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability.* 25. p. 287–298.

17. KOOPMAN, B. O. (1936) On distributions admitting a sufficient statistic. *Transactions of the American Mathematical society.* 39 (3). p. 399–409.

18. KULLBACK, S., LEIBLER, R. A. (1951) On information and sufficiency. *The annals of mathematical statistics.* 22 (1). p. 79–86.

19. BUBECK, S., CESA-BIANCHI, N. (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning.* 5 (1). p. 1–122.

20. SLIVKINS, A. (2019) Introduction to multi-armed bandits. *Foundations and Trends in Machine Learning.* 12 (1-2). p. 1–286.

21. DZHOHA, A. (2021) Bernoulli multi-armed bandit problem under delayed feedback. *Bulletin of Taras Shevchenko National University of Kyiv. Series: Physics & Mathematics.* №. 1. p. 20–26.

22. AUER, P., CESA-BIANCHI, N., FISCHER, P. (2002) Finite-time Analysis of the Multi-armed Bandit Problem. *Machine Learning.* 47 (2). p. 235–256.

23. HOEFFDING, W. (1963) Probability inequalities for sums of bounded random variables. *Journal of the American statistical association.* 58 (301). p. 13–30.

24. CHERNOFF, H. (1952) A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics.* 23 (4). p. 493–507.

25. GARIVIER, A., CAPPE, O. (2011) The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond. *Conference on Learning Theory, PMLR.* p. 359–376.

26. CAPPE O., GARIVIER, A., MAILLARD, O. A., MUNOS, R., STOLTZ, G. (2013) Kullback-Leibler Upper Confidence Bounds For Optimal Sequential Allocation. *Annals of Statistics.* 41 (3). p. 1516–1541.

27. AGRAWAL, S., GOYAL, N. (2012) Analysis of thompson sampling for the multi-armed bandit problem. *Conference on learning theory, JMLR.* p. 39.

28. BELLMAN, R. (1956) A problem in the sequential design of experiments. *The indian journal of statistics.* 16 (3/4). p. 221–229.

29. BERRY, D., FRISTEDT, B. (1985) Bandit Problems. Sequential allocation of experiments. *Chapman and Hall, Springer.* 5 (51-87). p. 7.

30. PUTERMAN, M. (1994) *Markov Decision Processes: discrete stochastic dynamic programming.* Wiley.

31. FELDMAN, D. (1962) Contributions to the "two-armed bandit". *The Annals of Mathematical Statistics.* 33 (3). p. 947–956.

32. BERRY, D. A. (1972) A Bernoulli two-armed bandit. *The Annals of Mathematical Statistics.* p. 871–897.

33. DZHOHA, A. (2021) *Multi-armed bandit problem under delayed feedback: numerical experiments.* [Online] Available from: https://github.com/djo/delayed-bandit.

Надійшла до редколегії 07.09.2021