Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Міністерство освіти і науки України
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Міністерство освіти і науки України

Кваліфікаційна наукова
праця на правах рукопису

КОЗИРЄВ АНТОН ЮРІЙОВИЧ

УДК 004.8:519.85

ДИСЕРТАЦІЯ
МЕТОДИ СТОХАСТИЧНОЇ КВАНТИЗАЦІЇ ДЛЯ МАСШТАБОВАНОГО КЛАСТЕРНОГО АНАЛІЗУ ВЕЛИКИХ ДАНИХ

Спеціальність 113 – Прикладна математика
Галузь знань 11 – Математика та статистика

Подається на здобуття наукового ступеня доктора філософії
Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

Козирєв Антон Юрійович

Науковий керівник: доктор фізико-математичних наук, старший науковий співробітник Норкін Володимир Іванович

Київ – 2026

АНОТАЦІЯ

Козирєв А.Ю. Методи стохастичної квантизації для масштабованого кластерного аналізу великих даних. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 113 «Прикладна математика». – Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ. – 2026.

Мета роботи – підвищення якості кластеризації та зменшення обчислювальних витрат під час оброблення великих даних, а також забезпечення стійкого до розподільного дрейфу ітеративного оновлення індексних структур у векторних базах даних без потреби в їх періодичній повній перебудові.

Для досягнення мети в роботі визначено наступні наукові завдання:

- аналіз існуючих методів кластеризації та виявлення їх переваг і недоліків при вирішенні задач кластерного аналізу;
- розробка методу стохастичної квазіградієнтної кластеризації на основі транспортної задачі з рухомими центрами за допомогою стохастичної апроксимації Кіфера-Вольфовіца;
- виведення умов збіжності запропонованого методу та дослідження швидкості збіжності з використанням елементів негладкого аналізу;
- запровадження структури даних стохастичного інвертованого файлового індексу на основі існуючої структури даних з використанням запропонованого методу для розбиття метричного простору вхідних даних;
- проведення експериментальних досліджень на відкритих наборах даних Iris, MNIST та ImageNet з оцінкою точності за допомогою статистичних метрик, таких як індекс Ранда та середньоквадратична помилка.

Наукова новизна. У дисертації одержано нові наукові результати:

1. Задачу кластеризації зведено до нової неопуклої негладкої задачі стохастичної оптимізації, що дозволяє для її розв'язання застосувати сучасні

адаптивні методи стохастичної оптимізації, які використовуються для навчання глибинних нейронних мереж;

2. Розроблено новий метод стохастичної квазіградієнтної кластеризації, який використовує субдиференційовану цільову функцію транспортної відстані та стохастичну апроксимацію, що дозволяє оцінювати оптимальні розміщення центроїдів лише з використанням підмножини навчальних даних і потребує меншу кількість обчислювальних ресурсів. Запропонований метод має теоретично доведені асимптотичні гарантії збіжності на основі властивостей узагальненої диференційованості цільової функції;

3. Враховуючи, що метод стохастичної квазіградієнтної кластеризації має доведену збіжність лише до локальних екстремумів цільової функції, запропоновано новий метод початкової ініціалізації центроїдів за допомогою жадібного алгоритму в комбінації з дискретною оптимізацією та модифікацію стохастичної квазіградієнтної кластеризації з використанням метаевристичного методу диференціальної еволюції, що дозволяє здійснювати пошук глобальних екстремумів негладкої неопуклої цільової функції та підвищити якість кластеризації за рахунок уникнення збіжності до неоптимальних локальних розв'язків;

4. Удосконалено скінченно-різницевий метод оптимізації негладких неопуклих цільових функцій за допомогою методу згладжування (усереднення) функцій для пошуку точок екстремумів, використовуючи скінченно-різницеву апроксимацію градієнтів вздовж стохастичних напрямків. Теоретично доведено сублінійну швидкість збіжності стохастичного скінченно-різницевого методу для Ліпшицевих опуклих функцій;

5. Удосконалено структуру стохастичного інвертованого файлового індексу для інкрементного індексування векторних даних, де для індексації використовується запропонований метод стохастичної квазіградієнтної кластеризації, в умовах динамічної зміни відповідного розподілу даних через сезонні зміни, соціологічні фактори, тощо.

Проблематика. Кластерний аналіз є основою сучасних систем штучного інтелекту, зокрема архітектур з пошуком за схожістю (RAG, RETRO) та векторних баз даних мільярдного масштабу, де інвертовані файлові індекси будуються

переважно методом K-середніх. Існуючі методи кластеризації на основі центроїдів мають низку суттєвих обмежень. По-перше, класичні методи K-середніх, K-медіан, гармонічних та нечітких K-середніх потребують обчислювальної складності та пам'яті порядку $O(Nd)$ на ітерацію, оскільки оновлення центроїдів вимагає повного проходу по всьому набору з N векторів розмірності d, що стає важкою задачею на масштабах N порядку 10^4-10^6 на типовому обчислювальному обладнанні. По-друге, цільова функція внутрішньокластерної суми квадратів є неопуклою та задача пошуку її глобального мінімуму NP-важка, а класичні детерміновані методи збігаються лише до локального оптимуму, який сильно залежить від початкової ініціалізації центроїдів. Навіть удосконалення як K-середніх++ забезпечують лише $O(\log K)$ гарантію. По-третє, після побудови індексу його центроїди та межі розбиття залишаються незмінними, що в умовах розподільного дрейфу призводить до зростання помилки квантизації, дисбалансу розмірів кластерів і деградації якості пошуку. Єдиним поширеним засобом протидії є періодична повна перебудова індексу, обчислювальна вартість якої при мільярдних масштабах сягає декількох діб. Крім того, цільова функція кластеризації $F(y) = \sum_i p_i \min_k \|\xi_i - y_k\|^r$ є негладкою через наявність функції мінімуму, що унеможливлює пряме застосування класичних градієнтних методів, а в задачах кластеризації даних високої розмірності проявляється «прокляття розмірності», за якого відстань втрачає здатність розрізняти об'єкти у вихідному просторі ознак.

Запропоноване вирішення. Для подолання зазначених проблем у роботі задачу оптимальної кластеризації переформульовано як транспортну задачу з рухомими центрами через відстань Канторовича-Рубінштейна (Вассерштейна) між емпіричним розподілом даних та дискретним розподілом, що його апроксимують K центроїдів. Аналітичне виключення транспортних змінних дозволяє звести вихідну задачу до неопуклої негладкої задачі стохастичної оптимізації $\min_y F(y) = \mathbb{E}_{i\sim p}[\min_k \|\xi_i - y_k\|^r]$. Застосування узагальненого диференціального числення до субдиференціала мінімуму дозволяє побудувати незміщений стохастичний субградієнт лише за одним випадковим спостереженням $\tilde{\xi}$ і визначити метод стохастичної квазіградієнтної кластеризації, у якому на кожній ітерації оновлюється лише центроїд, найближчий до вибраного зразка, що зменшує вимоги до пам'яті з $O(Nd)$ до $O(d)$ на крок. На цій основі розроблено адаптивні

варіанти, що адаптивно масштабують крок для кожного центроїда та автоматично балансують навантаження між часто й рідко оновленими кластерами. Замість імовірнісної ініціалізації K-середніх++ запропоновано детерміновану жадібну агрегацію близьких даних із розв'язанням задачі цілочисельного лінійного програмування з обмеженням на потужність множини центроїдів. Для отриманого методу аналітично доведено збіжність послідовності центроїдів до зв'язної компоненти множини стаціонарних точок за умов Роббінса-Монро на крок і компактності допустимої області. Структуру даних стохастичного інвертованого файлового індекса побудовано на основі еквівалентності задачі побудови розбиттів класичного індекса та задачі стохастичної квазіградієнтної кластеризації при $r = 2$ й рівномірних вагах: кожне додавання нового вектора супроводжується $O(d)$ стохастичним коригуванням найближчого центроїда з використанням адаптивної оцінки моментів, що забезпечує безперервну адаптацію індексу до зміни розподілу даних та сумісність з квантизацією добутку. Для подолання «прокляття розмірності» метод стохастичної квазіградієнтної кластеризації поєднано з трійковою нейронною мережею, яка через контрастне навчання відображає вихідні високовимірні дані в низьковимірний латентний простір, придатний для ефективної кластеризації.

Практичне значення отриманих результатів полягає в експериментальному підтвердженні збіжності запропонованого методу квантизації на відкритих наборах даних з порівнянням швидкості збіжності та точності оптимальних центрів. В експериментальних результатах показано, що запропонований метод в середньому скорочує кількість обчислень функції відстань приблизно на 70% порівняно з числом обчислень класичного методу кластерного аналізу K-середніх, який є стандартним методом кластеризації на основі центроїдів. Масштабованість методу продемонстрована на відкритих наборах даних з варіацією від 10^4 до 10^6 елементів.

Запропонований метод кластеризації може бути застосований для ітеративної індексації даних в дуже великих базах даних в умовах нестаціонарності. Також він може бути застосований для ітеративної апроксимації неперервних та дискретних ймовірнісних розподілів, для розв’язання задач логістики, розміщення складів, сервісних центрів.

Результати роботи впроваджено у навчальний процес кафедри штучного інтелекту ННІПСА Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського» в рамках освітнього компонента «Сучасні методи оптимізації» для здобувачів першого (бакалаврського) рівня вищої освіти за освітньо-професійною програмою «Системи та методи штучного інтелекту» (лекції 6, 10), що відображено в силабусі відповідної освітньої компоненти, ухваленому на засіданні кафедри штучного інтелекту (протокол №14 від 11.06.2024) та погодженому методичною комісією ННІПСА (протокол №10 від 24.06.2024), (url: https://ai.kpi.ua/ua/bachelors/syllabus/2024-2025/28343_3_suchasni_metody_optymizatsii.pdf).

Висновки. У дисертаційній роботі побудовано цілісну математичну та програмну основу для масштабованого кластерного аналізу великих даних, що ґрунтується на зведенні задачі оптимальної кластеризації до неопуклої негладкої задачі стохастичної оптимізації та застосуванні до неї сучасних адаптивних методів, перенесених із галузі глибинного навчання. Запропонований метод стохастичної квазіградієнтної кластеризації та його адаптивні модифікації забезпечують зменшення вимог до пам'яті з $O(Nd)$ до $O(d)$ на ітерацію та скорочення кількості обчислень функції відстані приблизно на 70% порівняно з класичним методом K-середніх при збереженні точності квантизації, а також збіжність методу до стаціонарної точки доведено аналітично за умов Роббінса-Монро. Розроблена структура даних стохастичного інвертованого файлового індексу вперше надає теоретично обґрунтований механізм інкрементного, стійкого до розподільного дрейфу обслуговування інвертованих файлових індексів у векторних базах даних, що усуває потребу в періодичній повній перебудові індексу та інтегрується з існуючими системами наближеного пошуку (зокрема FAISS) і продуктовою квантизацією. Поєднання методу стохастичної квазіградієнтної кластеризації з контрастним навчанням трійкових нейронних мереж дозволяє подолати «прокляття розмірності» та досягти на наборі MNIST якості кластеризації, порівнянної з повністю керованими базовими методами, зберігаючи стійкість в умовах обмеженого розмічування даних. Практична значущість роботи підтверджена застосуваннями до стиснення кольорових зображень з втратами, потокової кластеризації в умовах сезонних та соціологічних змін розподілу даних, а

також інтеграції з системами векторного пошуку та підтримки прийняття рішень; результати впроваджено в навчальний процес кафедри штучного інтелекту ННІПСА КПІ ім. Ігоря Сікорського. Перспективними напрямами подальших досліджень є поширення розробленого стохастичного підходу на моделі гаусівських сумішей, методи кластеризації на основі щільності тощо.

Ключові слова: теорія оптимізації, стохастична оптимізація, негладка оптимізація, адаптивні градієнтні методи, кластерний аналіз, векторна квантизація, K-середніх, машинне навчання, глибинне навчання, аналіз даних, статистика, штучний інтелект, інтелектуальний аналіз тексту, семантична нейронна мережа.

ANNOTATION

Kozyriev A.Yu. Methods of Stochastic Quantization for Scalable Cluster Analysis of Big Data. – Qualifying scientific work as a manuscript.

Thesis for the degree of Doctor of Philosophy in the specialty 113 «Applied Mathematics». – National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute», Kyiv. – 2026.

The purpose of the work is to improve the quality of clustering and reduce computational costs when processing big data, as well as to ensure iterative update of index structures in vector databases that is robust to distributional drift without the need for their periodic complete reconstruction.

To achieve the stated purpose, the following research tasks have been defined:

- conduct an analysis of existing clustering methods and identify their advantages and disadvantages in solving cluster analysis problems;
- develop the stochastic quasigradient clustering method based on the transportation problem with moving centers using the Kiefer–Wolfowitz stochastic approximation;
- derive convergence conditions of the proposed method and investigate the convergence rate using elements of non-smooth analysis;
- introduce a stochastic inverted file index data structure based on an existing data structure using the proposed method to partition the metric space of the input data;
- conduct experimental studies on open datasets Iris, MNIST, and ImageNet and assess the accuracy of the obtained results using statistical metrics such as the Rand index and mean squared error.

Scientific novelty. The thesis presents the following novel scientific results:

1. The clustering problem is reduced to a new non-convex non-smooth stochastic optimization problem, which allows for its solution to apply modern adaptive stochastic optimization methods used for training deep neural networks;
2. A new stochastic quasigradient clustering method is developed, which uses a subdifferentiated transport distance objective function and stochastic approximation, which allows estimating optimal centroid placements using only a subset of the training

data and requires fewer computational resources. The proposed method has theoretically proven asymptotic convergence guarantees based on the properties of generalized differentiability of the objective function;

3. Considering that the stochastic quasi-gradient clustering method has proven convergence only to local extrema of the objective function, a new method of initial centroid initialization using a greedy algorithm in combination with discrete optimization and a modification of stochastic quasigradient clustering using the metaheuristic method of differential evolution is proposed, which allows searching for global extrema of a non-smooth non-convex objective function and improving the quality of clustering by avoiding convergence to suboptimal local solutions;

4. The finite-difference method of optimizing non-smooth non-convex objective functions is improved using the method of smoothing (averaging) functions to search for extrema points, using finite-difference approximation of gradients along stochastic directions. The sublinear convergence rate of the stochastic finite-difference method for Lipschitz convex functions has been theoretically proven;

5. The structure of the stochastic inverted file index for incremental indexing of vector data has been improved, where the proposed stochastic quasi-gradient clustering method is used for indexing, under conditions of dynamic change of the corresponding data distribution due to seasonal changes, sociological factors, etc.

Problem statement. Cluster analysis constitutes the foundation of modern artificial intelligence systems, particularly similarity-search architectures (RAG, RETRO) and billion-scale vector databases, where inverted file indices are constructed predominantly by means of the K-means method. Existing centroid-based clustering methods exhibit a number of substantial limitations. First, the classical K-means, K-medians, harmonic and fuzzy K-means methods exhibit per-iteration computational and memory complexity of order $O(Nd)$, since updating the centroids requires a full pass over the entire set of N vectors of dimension d, which becomes infeasible on commodity hardware as N reaches 10^4-10^6. Second, the within-cluster sum-of-squares objective is non-convex, finding its global optimum is NP-hard, and deterministic methods converge only to a local optimum whose quality depends strongly on the initial choice of centroids; even refinements such as K-means++ provide only an $O(\log K)$ guarantee. Third, once an stochastic inverted file index has been built, its centroids and partition boundaries remain

frozen, so that under dynamic distributional drift (seasonal changes, sociological factors, evolving content) the quantization error grows, the partition sizes become unbalanced, and the quality of downstream retrieval and generative models deteriorates; the only commonly used remedy is periodic full re-indexing, whose computational cost at billion-vector scale amounts to several days. In addition, the centroid-based clustering objective $F(y) = \sum_i p_i \min_k \|\xi_i - y_k\|^r$ is non-smooth because of the minimum function, which precludes direct application of classical gradient methods, while high-dimensional clustering is further impeded by the curse of dimensionality, under which distance loses its discriminative power in the raw feature space.

Proposed solution. To overcome the above problems, the optimal clustering problem is reformulated as a transportation problem with movable centers based on the Kantorovich-Rubinstein (Wasserstein) distance between the empirical data distribution and the discrete distribution induced by K centroids. Analytical elimination of the transport variables reduces the original problem to a non-convex non-smooth stochastic optimization problem $\min_y F(y) = \mathbb{E}_{i\sim p}[\min_k \|\xi_i - y_k\|^r]$. Applying the generalized differentiation to the subdifferential of the minimum function makes it possible to construct an unbiased stochastic subgradient from a single random sample $\tilde{\xi}$ and to define the stochastic quasigradient clustering method, in which at every iteration only the centroid nearest to the drawn sample is updated; this reduces the per-step memory requirement from $O(Nd)$ to $O(d)$. On this basis, adaptive variants have been developed, which adaptively rescale the step size for each centroid and automatically balance the load between frequently and rarely updated clusters. Instead of the probabilistic K-means++ initialization, a deterministic greedy aggregation of nearby data combined with a cardinality-constrained integer linear program over the candidate centers has been proposed. For the resulting method, almost-sure convergence of the centroid sequence to a connected component of the set of stationary points has been proved analytically under Robbins-Monro conditions on the step size and compactness of the feasible region; the result has been extended to the adaptive variant. The stochastic inverted file index data structure is constructed by exploiting the equivalence between inverted file index codebook training and the stochastic quasigradient clustering objective with $r = 2$ and uniform weights: each insertion of a new vector is accompanied by an $O(d)$ stochastic update of the nearest centroid using the adaptive moments, which provides continuous

adaptation of the index to data-distribution drift and compatibility with product quantization. To overcome the curse of dimensionality, the stochastic quasigradient clustering method is composed with a triplet neural network that, via contrastive learning, maps the original high-dimensional data into a low-dimensional latent space suitable for effective clustering.

The practical significance of the results obtained lies in the experimental confirmation of convergence of the proposed quantization method on open datasets with comparison of convergence speed and accuracy of optimal centers. Experimental results show that, on average, the proposed method reduces the number of distance-function evaluations by approximately 70% compared to the number of evaluations performed by the classical K-means cluster analysis method, which is the standard centroid-based clustering method. The scalability of the method has been demonstrated on open datasets ranging in size from 10^4 to 10^6 elements.

The proposed clustering method can be applied to iterative data indexing in very large databases under conditions of data changes. It can also be applied to iterative approximation of continuous and discrete probability distributions, and to solving problems of logistics, warehouse placement, and service-center location.

The results of the work have been incorporated into the educational process of the Department of Artificial Intelligence of the Educational and Scientific Institute for Applied System Analysis (ESIASA) of the National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute» within the educational component «Modern Optimization Methods» for first (bachelor's) level higher education applicants of the educational-professional program «Systems and Methods of Artificial Intelligence» (lectures 6, 10), which is reflected in the syllabus of the corresponding educational component, approved at the meeting of the Department of Artificial Intelligence (protocol No. 14 of 11.06.2024) and by the methodological commission of ESIASA (protocol No. 10 of 24.06.2024), (url: https://ai.kpi.ua/ua/bachelors/syllabus/2024-2025/28343_3_suchasni_metody_optymizatsii.pdf).

Conclusions. The thesis provides a unified mathematical and software foundation for scalable centroid-based cluster analysis of big data, based on the reduction of the optimal clustering problem to a non-convex non-smooth stochastic optimization problem and the application to it of modern adaptive methods transferred from the field of deep

learning. The proposed stochastic quasigradient clustering method and its adaptive modifications reduce the per-iteration memory requirement from $O(Nd)$ to $O(d)$ and lower the number of distance-function evaluations by approximately 70% with respect to the classical K-means method while preserving quantization accuracy; almost-sure convergence of the method to a stationary point has been proved analytically under Robbins-Monro conditions. The developed stochastic inverted file index data structure provides, for the first time, a theoretically grounded mechanism for incremental, drift-resilient maintenance of inverted file indices in vector databases, eliminating the need for periodic full re-indexing and integrating seamlessly with existing approximate-nearest-neighbor systems (such as FAISS) and with product quantization. The combination of stochastic quasigradient clustering with contrastive triplet-network learning has made it possible to overcome the «curse of dimensionality» and to attain MNIST clustering quality comparable with fully supervised baselines, while remaining robust under limited data labeling. The practical significance of the work has been confirmed by applications to lossy color image compression, streaming clustering under seasonal and sociological distribution changes, and integration with vector-search and decision-support systems; the results have been incorporated into the educational process of the Department of Artificial Intelligence of ESIASA at the Igor Sikorsky Kyiv Polytechnic Institute. Promising directions for further research include the extension of the developed stochastic approach to Gaussian mixture models, density-based clustering methods etc.

Keywords: optimization theory, stochastic optimization, non-smooth optimization, adaptive gradient methods, cluster analysis, vector quantization, K-means method, machine learning, deep learning, data analysis, statistics, artificial intelligence, text mining, semantic network.

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Статті у періодичних наукових виданнях, проіндексованих у базах даних Web of Science Core Collection та/або Scopus

1. Norkin, V., Pichler, A., & Kozyriev, A. (2025). Constrained Global Optimization by smoothing. Lecture Notes in Computer Science, Vol. 14476. P.136–150. Print ISSN 0302-9743.

   У роботі здобувач приймав участь у підготовці огляду літератури, отриманні оцінок швидкості збіжності методу, розробив алгоритм і програмну реалізацію, провів тестування алгоритму на тестових прикладах, сформулював загальні висновки та інтерпретацію отриманих результатів. Співавтор Norkin, V. виконав теоретичну частину з отримання оцінки швидкості збіжності методу та з доведення теореми про збіжність, сформулював основну частину технічного звіту у публікації та керував виконанням дослідження. Співавтор Pichler, A. провів огляд літератури та аналіз існуючих методів, а також сформулював порівняльну характеристику з запропонованим методом.

Статті у наукових виданнях, включених на дату опублікування до переліку наукових фахових видань України за спеціальністю 113

2. Norkin, V., Kozyriev, A. and Norkin, B. (2024) “Modern stochastic quasi-gradient optimization algorithms”, International Scientific Technical Journal "Problems of Control and Informatics", 69(2), pp. 71–83.

   У роботі здобувач зробив огляд та аналіз літератури, провів тестування стохастичних адаптивних скінченно-різницевих методів, провів статистичну обробку даних і порівняльний аналіз результатів. Співавтор Norkin, V. виконав детальний аналіз стохастичних градієнтних методів та їх властивостей, а також керував виконанням дослідження. Співавтор Norkin, B. створив програмне забезпечення для порівняння характеристик розглянутих скінченно-різницевих методів.

3. Kozyriev, A., & Norkin, V. (2024). Robust Clustering on High-Dimensional Data with Stochastic Quantization. International Scientific Technical Journal “Problems of Control and Informatics,” 70(1), pp. 32–48.

   У роботі здобувач зробив огляд літератури, запропонував архітектуру нейронної мережі для кодування зображень в простір латентних ознак, програмно реалізував метод стохастичної квантизації, протестував підхід на тестових наборах даних. Співавтор Norkin, V. запропонував метод кластеризації на основі стохастичної апроксимації, сформулював процедуру зведення транспортної задачі з рухомими центрами у задачу неопуклої негладкої оптимізації, провів аналіз умов збіжності запропонованого метода відносно заданих значень гіперпараметрів та керував виконанням дослідження.

4. Kozyriev, A., & Norkin, V. (2024). Lossy image compression with stochastic quantization. Cybernetics and Computer Technologies, (3), 60–66.

   У роботі здобувач зробив огляд літератури, провів аналіз наявних підходів до вирішення задачі стиснення кольорових зображень, запропонував тестову задачу кольорової квантизації, реалізував алгоритм, зробив ілюстрації. Співавтор Norkin, V. виконав огляд алгоритмів кодування та зберігання зображень на цифрових пристроях, а також керував виконанням дослідження.

5. Norkin, V., & Kozyriev, A. (2023). On Shor’s R-algorithm for problems with constraints. Cybernetics and Computer Technologies, (3), 16–22.

   У роботі здобувач реалізував алгоритм Шора та провів тестові розрахунки, зробив кінцеве редагування публікації.Співавтор Norkin, V. зробив огляд літератури за темою алгоритму Шора, сфомулював теоретичну частину публікації та керував виконанням дослідження.

Матеріали конференцій

6. Norkin V., Pichler A., Kozyriev A. Constrained global optimization by smoothing. Book of Abstracts of the 4th International Conference and Summer School “Numerical Computations: Theory and Algorithms (NUMTA 2023)” (Calabria,

Italy, June 14-20, 2023), edited by Y.D. Sergeyev, D.E. Kvasov, M. Chiara Nasso. P.59.

У роботі здобувач зробив огляд літератури. Співавтор Norkin, V. виконав теоретичну частину з отримання оцінки швидкості збіжності методу та з доведення теореми про збіжність, сформулював основну частину технічного звіту. Співавтор Pichler, A. сформулював порівняльну характеристику з запропонованим методом.

7. Козирєв А.Ю., Норкін В.І. Стиснення зображення з втратами за допомогою стохастичного квантування. Збірник тез доповідей XXVII Міжнародної конференції «АВТОМАТИКА 2024», присвяченої пам’яті академіка НАН України В.М. Кунцевича та академіка НАН України Ю.Г. Кривоноса. 20–22 листопада 2024 р., м. Дніпро (Україна) . ISBN 978-966-2344-74-5 . С.124-125.
У роботі здобувач зробив огляд літератури. Співавтор Norkin, V. виконав огляд алгоритмів кодування та зберігання зображень на цифрових пристроях, а також керував виконанням дослідження.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

AdaGrad – adaptive gradient (метод адаптивного градієнта);

ADAM – adaptive moment estimation (метод адаптивної оцінки моментів);

ANN – approximate nearest neighbors (наближений пошук найближчих сусідів);

DNN – deep neural network (глибинна нейронна мережа);

DE – differential evolution (метод диференціальної еволюції);

DL – deep learning (глибинне навчання);

EM – Expectation-Maximization;

FAISS – Facebook AI Similarity Search (пошук найближчих сусідів за допомогою ШІ у Facebook);

GMM – Gaussian mixture model (модель гаусівської суміші);

IVF – inverted file index (інвертований файловий індекс);

IVFADC – Inverted File with Asymmetric Distance Computation;

JPEG – Joint Photographic Experts Group;

MSE – mean squared error (середньоквадратичну помилку);

NLP – natural language processing (задача машинної обробки тексту);

PCA – principal component analysis (метод головних компонент);

PQ – product quantization (квантизація добутку);

RAG – retrieval augmented generation (генерація тексту з доповненням через пошук);

RETRO – retrieval-enhanced transformer (архітектура мережі Трансформер з доповненням через пошук);

RMSProp – root-mean squared propagation (метод середньоквадратичного спуску);

SIVF – stochastic inverted file index (стохастичний інвертований файловий індекс);

SQC – stochastic quasigradient clustering (метод стохастичної квазіградієнтної кластеризації);

АЗРП – алгоритм зворотного розповсюдження помилки (backpropagation);

ІАТ – інтелектуальний аналіз тексту (text mining);

ЗЦЛП – змішане цілочисельне лінійне програмування;

ШІ – штучний інтелект.

ВСТУП

Актуальність теми дослідження. Кластерний аналіз є класичною задачею в галузі аналізу даних, яка є основою методології для розуміння внутрішньої структури складних наборів даних. Задача організації об'єктів у змістовні групи на основі їхньої подібності відігравало незамінну роль протягом усієї історії розвитку людських знань [129]. Кластерний аналіз перетворився з концептуальної основи для організації спостережень на складний набір обчислювальних методів, що лежать в основі різних прикладних задач, починаючи від генетики [53] та сегментації ринку [120] і закінчуючи обробкою зображень [49] та виявленням аномалій [20]. Формалізацію кластерного аналізу як обчислювальної задачі можна простежити до середини XX століття [70], коли експоненціальне зростання можливостей збору даних вимагало підходів до організації та інтерпретації даних. Основний принцип, що лежить в основі кластерного аналізу, залишається концептуально простим: маючи набір об'єктів, що характеризуються вимірюваними атрибутами, розділити ці об'єкти на підмножини таким чином, щоб об'єкти в одній підмножині мали більшу схожість один з одним, ніж з об'єктами в різних підмножин [129].

Галузь кластерного аналізу останнім часом отримала більше уваги через сплеск розробки та досліджень штучного інтелекту в останні роки. Згідно річного звіту 2025 року «індексу штучного інтелекту» Стенфордського університету [75], приватний капітал, що надходить у компанії, що займаються штучним інтелектом, досяг приблизно 252.3 мільярда доларів США у світі – охоплюючи весь «стек» штучного інтелекту від інфраструктури до лабораторій та застосувань на основі базових моделей – і становив на 25.5% більше інвестування за попередній рік. Такий ріст фінансування пов'язаний з нещодавніми досягненнями в задачах машинної обробки тексту, які призвели до появи архітектур із зовнішнім механізмом пам'яті, що розширюють можливості генеративних моделей шляхом інтеграції із зовнішніми базами знань. Серед цих розробок слід відзначити системи RETRO [13] та RAG [68], які використовують метод наближеного пошуку найближчих сусідів для динамічного отримання контекстуально релевантної

інформації з великомасштабних сховищ документів під час генерації відповіді мовною моделлю [49].

У цих архітектурах метод наближеного пошуку найближчих сусідів служить механізмом для ідентифікації та отримання найбільш семантично релевантних векторів з масивних зовнішніх сховищ пам'яті, що дозволяє генеративній моделі ґрунтувати свої вихідні дані на отриманому контенті, а не покладатися виключно на параметричні знання, отримані під час навчання [13]. Для досягнення обчислювальної придатності в масштабі ці системи пошуку зазвичай використовують структури інвертованого файлового індексу, які розділяють векторний простір на непересічні розбиття шляхом кластеризації, зазвичай з використанням K-середніх, тим самим обмежуючи пошук лише тими розділами, які, найімовірніше, містять відповідних сусідів [7,49]. Ця стратегія розділення простору значно зменшує складність пошуку, уникаючи зайвих обчислень відстаней по всій базі даних.

Однак, залежність від інвертованого файлового індексу вводить обмеження в динамічних середовищах: структура індексу залишається стаціонарною після побудови, зберігаючи фіксовані центроїди кластера та межі розділів, навіть коли базовий розподіл даних розвивається шляхом постійного додавання нових векторів [7]. Цей феномен, відомий як розподільний дрейф, при якому статистичні властивості вхідних даних відрізняються від властивостей навчального набору, що використовується для побудови індексу, поступово погіршує якість пошуку та ефективність пошуку, проявляючись у збільшенні помилок квантизації, незбалансованих розмірах розділів та, зрештою, далі погіршує продуктивність моделі ШІ, яка залежить від точного пошуку для контекстного обґрунтування. Єдиний існуюче рішення проблеми розподільного дрейфу на даний момент є періодична повторна індексація усієї множини даних, що потребує значних обчислювальних ресурсів з ростом кількості елементів в базі даних.

Зв’язок роботи з науковими програмами, планами, темами.. Дисертаційна робота виконана відповідно до науково-технічної діяльності Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського» і кафедри прикладної математики.

Мета і завдання дослідження. Метою роботи є підвищення якості кластеризації та зменшення обчислювальних витрат під час оброблення великих даних, а також забезпечення стійкого до розподільного дрейфу ітеративного оновлення індексних структур у векторних базах даних без потреби в їх періодичній повній перебудові.

Для досягнення поставленої мети в рамках дисертації були визначені наступні наукові завдання:

1. Провести аналіз існуючих методів векторної квантизації, виявити їхні переваги та недоліки при вирішенні задач кластерного аналізу;
2. Розробити метод SQC за основі транспортної задачі з рухомими центрами [64] за допомогою стохастичної апроксимації Кіфера-Вольфовіца [55];
3. Вивести умови збіжності запропонованого методу використовуючи елементи негладкого аналізу;
4. Запровадити структуру даних SIVF на основі існуючої структури даних IVF з використанням запропонованого методу для розбиття метричного простору вхідних даних;
5. Провести експериментальні дослідження на відкритих даних Iris [34], MNIST [67] та ImageNet [108] і оцінити точність отриманих результатів, використовуючи статистичні метрики як індекс Ранда [102] та середньоквадратичну помилку.

Об'єкт дослідження. Об'єктом дослідження є методи стохастичної оптимізації, субградієнтні методи, скінченно-різницеві методи, методи машинного навчання «без учителя», методи кластеризації на основі центроїдів, транспортні задачі з рухомими пунктами призначення.

Предмет дослідження. Предметом дослідження є методи стохастичної квазіградієнтної кластеризації та модифікацій з адаптивним кроком. Показники оцінки точності та швидкості збіжності методів стохастичної оптимізації.

Методи дослідження. У рамках дослідження використовувалися теорія стохастичної оптимізації, елементи негладкого аналізу та методи глибинного навчання, зокрема нейронні мережі представлень латентного простору для перетворення дискретних даних у неперервні векторні репрезентації в метричному просторі. Експериментальні дослідження проводилися на даних у відкритому

доступі з різною розмірністю вхідних ознак, зокрема Iris [34] (низьковимірний набір даних про виміри квітів ірису), MNIST [67] (зображення 28×28 рукописних цифр у відтінках сірого) та ImageNet [108] (повнокольорові зображення, приведені до уніфікованої роздільної здатності 128×128 пікселів для прикладної задачі кольорової квантизації).

Наукова новизна отриманих результатів. У дисертації одержано нові наукові результати:

1. Задачу кластеризації зведено до нової неопуклої негладкої задачі стохастичної оптимізації, що дозволяє для її розв'язання застосувати сучасні адаптивні методи стохастичної оптимізації, які використовуються для навчання глибинних нейронних мереж;

2. Розроблено новий метод стохастичної квазіградієнтної кластеризації, який використовує субдиференційовану цільову функцію транспортної відстані та стохастичну апроксимацію, що дозволяє оцінювати оптимальні розміщення центроїдів лише з використанням підмножини навчальних даних і потребує меншу кількість обчислювальних ресурсів. Запропонований метод має теоретично доведені асимптотичні гарантії збіжності на основі властивостей узагальненої диференційованості цільової функції;

3. Враховуючи, що метод стохастичної квазіградієнтної кластеризації має доведену збіжність лише до локальних екстремумів цільової функції, запропоновано новий метод початкової ініціалізації центроїдів за допомогою жадібного алгоритму в комбінації з дискретною оптимізацією та модифікацію стохастичної квазіградієнтної кластеризації з використанням метаевристичного методу диференціальної еволюції, що дозволяє здійснювати пошук глобальних екстремумів негладкої неопуклої цільової функції та підвищити якість кластеризації за рахунок уникнення збіжності до неоптимальних локальних розв'язків;

4. Удосконалено скінченно-різницевий метод оптимізації негладких неопуклих цільових функцій за допомогою методу згладжування (усереднення) функцій для пошуку точок екстремумів, використовуючи скінченно-різницеву апроксимацію градієнтів вздовж стохастичних напрямків. Теоретично доведено

сублінійну швидкість збіжності стохастичного скінченно-різницевого методу для Ліпшицевих опуклих функцій;

5. Удосконалено структуру стохастичного інвертованого файлового індексу для інкрементного індексування векторних даних, де для індексації використовується запропонований метод стохастичної квазіградієнтної кластеризації, в умовах динамічної зміни відповідного розподілу даних через сезонні зміни, соціологічні фактори, тощо.

Наукове та практичне значення роботи. Розроблені у дисертації алгоритм та структура даних мають значне наукове значення, оскільки забезпечують адаптивне оновлення індексу IVF векторних баз даних в реальному часі з мінімальними затратами на обчислювальні ресурси та високою точністю семантичного пошуку. Структура даних може бути застосована в довільних задачах з використанням векторного пошуку, наприклад, в пошукових рушіях тексту, зображень тощо. Запропонований алгоритм також розширює можливості кластерного аналізу в умовах потоків даних, де дані безперервно збираються та індексуються протягом тривалих періодів часу, а статистичні властивості часто змінюються через сезонні коливання, соціологічні тенденції тощо.

Практичне значення роботи полягає у потенціалі використання розробленої структури даних в системах керування векторними базами даних та індексів, наприклад FAISS [48], для більш ефективного та точного пошуку високо-розмірних медіаданих за семантичною схожістю. Запропонована структура даних SIVF може бути інтегрована у сучасні пошукові рушії та системи підтримки прийняття рішень, основаних на великих даних, для більш точного кластерного аналізу та виявлення закономірностей на нерозмічених даних.

Науково-практичну важливість результатів дисертації підтверджує впровадження результатів досліджень у навчальний процес на факультеті прикладної математики та навчально-наукового інституту прикладного системного аналізу Національного Технічного Університету України «Київського політехнічного інституту імені Ігоря Сікорського». Елементи розроблених методів та підходів включені до курсу «Сучасні методи оптимізації» в рамках аспірантської асистентської практики. Це дозволяє студентам ознайомитися з сучасними методами кластерного аналізу основаних на теорії стохастичної оптимізації та

елементах негладкого аналізу, підвищуючи ефективність навчального процесу та рівень підготовки фахівців.

Таким чином, результати дисертаційного дослідження мають як наукове значення для подальшого розвитку методів кластерного аналізу та стохастичної оптимізації, так і практичне значення, підтверджене використанням у пакетах програмного забезпечення і освітньому процесі.

Особистий внесок здобувача. У роботах, які виконано у співавторстві, на захист виносяться виключно результати, отримані особисто здобувачем, зокрема було здійсненно:

1. аналіз наявних підходів до вирішення поставленої задачі;
2. формулювання та обґрунтування наукової гіпотези;
3. розробку запропонованої структури даних стохастичного IVF;
4. проведення експериментів, статистичну обробку даних і порівняльний аналіз результатів;
5. формулювання загальних висновків та інтерпретацію отриманих результатів.

Співавторам у відповідних працях належали постановка окремих підзадач, консультацій щодо теоретичного обґрунтування отриманих результатів та технічна підтримка в процесі створення програмного забезпечення для експериментального дослідження запропонованої структури даних.

Апробація результатів дисертації. Результати дисертації доповідались на:

1. IV Міжнародна конференція «Чисельні обчислення: теорія та алгоритми (NUMTA 2023)», м. Калабрія, Італія, 14-20 червня 2023 р. [85];
2. XVI Міжнародна конференція «Освіта та дослідження в інформаційному суспільстві (ERIS 2023)», м. Пловдів, Болгарія, 12 жовтня 2023 р. [89];
3. XXVII Міжнародна конференція «АВТОМАТИКА 2024», присвяченої пам’яті академіка НАН України В.М. Кунцевича та академіка НАН України Ю.Г. Кривоноса, м. Дніпро, 20–22 листопада 2024 р. [57];
4. XVII Міжнародна конференція «Освіта та дослідження в інформаційному суспільстві (ERIS 2025)», м. Пловдів, Болгарія, 3-4 листопада 2025 р. [91];

5. XIV Міжнародна науково-практична конференція «Сучасна кібернетика: Глушковські читання-2025» Інституту кібернетики імені В.М. Глушкова НАН України, м. Київ, 26 лютого 2024 р. [59]

Публікації. Основні наукові результати дисертаційної роботи викладено в 5 публікаціях, з яких: 4 статті опубліковано в фахових журналах України, 1 стаття опублікована англійською мовою в науковому журналі видавництва Springer та проіндексована в наукометричній базі SCOPUS, 2 тези доповідей опубліковано в збірниках доповідей міжнародних та всеукраїнських наукових конференцій.

Структура та обсяг дисертації.

В **першому розділі** зроблено огляд існуючих методів кластерного аналізу та векторної квантизації. Сформульовано постановку задачі машинного навчання «без учителя» та наведено приклади прикладних задач, що потребують застосування кластерного аналізу. Розглянуто метод K-середніх та його похідні (K-медіан, гармонічних K-середніх, нечіткі модифікації), з акцентом на умовах збіжності, стратегіях ініціалізації, алгоритмічній складності та стійкості до статистичних шумів. Окрему увагу приділено неопуклості цільової функції методів кластеризації та основній проблемі використання сучасних методів векторної квантизації в умовах потоків даних, а саме проблемі розподільного дрейфу.

В **другому розділі** розглядаються адаптивні стохастичні градієнтні методи, що використовуються в запропонованому методі SQC для ітеративної оцінки оптимальних значень центроїдів елементів множини вхідних даних. Розглянуто властивості кожного з запропонованих методів та порівняно швидкість збіжності за заданих початкових умов. Також показано застосування стохастичної апроксимації до розглянутих градієнтних методів для зменшення навантаження на пам'ять обчислювального пристрою та контролю дисперсії оцінки градієнтів.

В **третьому розділі** розглянуто властивості узагальнено-диференційованих функцій та методів для знаходження оптимальних значень для негладкої цільової функції, що дозволяє застосування розглянутих в попередніх розділах методів до запропонованого методу SQC. Запропоновано новий скінченно-різницевий метод для оптимізації негладких цільових функції, використовуючи згладжування за Стєкловим. Теоретично доведено збіжність запропонованого методу та показано швидкість збіжності як для фіксованого кроку спуску так і для змінного кроку в

часі. Окремо розглянуто метод Шора як представник класичних узагальнених градієнтних методів для негладкої оптимізації.

В **четвертому розділі** описано новий запропонований метод SQC для векторної квантизації множини даних з метричного простору, виведеного на основі транспортної задачі з рухомими центрами. Показано процес зведення задачі до стохастичної форми використовуючи метод апроксимації з попередніх розділів, а також побудовано адаптивні модифікації методу на основі AdaGrad, RMSProp та ADAM. Розглянуто запропонований детермінований метод ініціалізації початкових центроїдів для методу SQC, що поєднує жадібний алгоритм попереднього відбору кандидатів із задачею ЗЦЛП з обмеженням кардинальності, та проведено порівняння зі стратегією ініціалізації K-середніх++. Теоретично доведено за яких умов запропонований метод є збіжним до зв'язної компоненти множини стаціонарних точок з ймовірністю один. Також розглянуто застосування методів глобальної оптимізації до неопуклої цільової функції квантизації.

В **п'ятому розділі** продемонстровано прикладне застосування запропонованого методу SQC та його адаптивних модифікацій до трьох суттєво відмінних класів задач. Спочатку розглянуто задачу кольорової квантизації зображень з порівняльними експериментами на підмножині набору даних ImageNet. Далі досліджено інтеграцію методу SQC з контрастивним навчанням за допомогою архітектури Triplet Network для розв'язання задач класифікації з підмножиною розмічених високовимірних даних, з оцінкою точності кластеризації на наборі даних MNIST за допомогою індексу Ранда. Зрештою, представлено застосування методу SQC для ітеративного оновлення центроїдів інвертованого файлового індексу (SIVF), що забезпечує поступову адаптацію структури розбиття до змін у розподілі векторів без потреби у повній реконструкції індексу.

Дисертація містить: анотацію, вступ, 5 розділів, висновки, список використаних джерел. Обсяг дисертації – 177 сторінок, основної частини – 146 сторінок. Робота містить 10 таблиць, 14 рисунків, список використаних джерел з 133 найменувань.

РОЗДІЛ 1. ОГЛЯД МЕТОДІВ КЛАСТЕРНОГО АНАЛІЗУ

У цьому розділі здійснено огляд наявних методів кластерного аналізу, що використовуються для знаходження розбиття точок на кластери, яке оптимізує явний критерій (наприклад, відстань в метричному просторі). Спочатку розглянуто постановку задачі машинного навчання «без учителя». Показано основні характеристики даного класу задач та приклади застосування методів на реальних задачах.

Далі було виконано огляд методів векторної квантизації – окремий клас методів кластеризації на основі центроїдів. Розглянуто метод K-середніх та похідні методи K-медіан, гармонічних K-середніх та модифікований метод для нечітких множин. Під час огляду було зроблено акцент на умовах збіжності методів, стратегіях ініціалізації початкових наближень алгоритмічної складності, та стійкості до статистичних шумів. Особливу увагу приділено проблемі неадаптивності отриманих центроїдів та подальшій проблемі розподільного дрейфу в динамічних середовищах коли кількість даних для аналізу та їх відповідний розподіл змінюється з часом.

Зрештою, сформульовано постановку задачі машинного навчання «без учителя» та показано роль методів векторної квантизації в задачах кластерного аналізу, їхні переваги та недоліки. Акцентована проблема неадаптивності до розподільного дрейфу є загальною характеристикою розглянутих існуючих методів, яка мотивує дане дослідження, метою якого є розробка альтернативного алгоритму, що демонструє стійкість до часових та контекстних змін у статистичних властивостях процесу агрегацій та аналізу даних.

1.1 Приклади задач кластерного аналізу

Приклад: Пошук текстової інформації у векторних базах даних. У роботі розглядається задача пошуку релевантної інформації у великомасштабних корпусах текстових даних, до яких можуть належати наукові архіви (математичні, фізичні, хімічні), а також контент соціальних мереж. Класичні підходи до

інформаційного пошуку базуються на індексації текстів за словами та подальшому зіставленні запиту з індексованими структурами. Сучасні підходи передбачають відображення текстів у векторний простір ознак. Семантична близькість текстів у цьому просторі визначається за допомогою метрик, зокрема евклідової або косинусної. Остання є більш адекватною у випадках, коли тексти відрізняються за довжиною, але мають подібну структуру частот слів. Задача пошуку за змістом формулюється як задача знаходження найближчих сусідів у векторному просторі: запит також відображається у вигляді вектора, після чого здійснюється пошук точок, найближчих до нього за обраною метрикою. Оскільки прямий перебір всіх точок є обчислювально неефективним для великих масивів даних, застосовуються методи попередньої кластеризації. Дані групуються у кластери навколо центрів (центроїдів), після чого пошук виконується у два етапи: спочатку визначаються найближчі центроїди, а далі – найближчі точки всередині відповідних кластерів.

Приклад: Оптимальне розміщення шкіл у сільській місцевості. Розглянемо сільськогосподарський регіон, що складається з великої кількості сіл і хуторів, у яких проживають сім’ї з дітьми. Діти повинні щоденно відвідувати школу протягом багатьох років. Очевидно, що побудувати школу в кожному населеному пункті неможливо, тому необхідно обрати обмежену кількість пунктів, у яких ці школи буде розміщено. З інших населених пунктів дітей доведеться організовано підвозити до шкіл щодня. Школи доцільно розміщувати у більш великих населених пунктах, однак навіть у цьому випадку їх кількість буде значно меншою за кількість усіх поселень. Таким чином, виникає задача вибору найкращого розміщення шкіл. У загальному випадку це комбінаторна задача, що передбачає перебір можливих варіантів розміщення. Для кожного варіанта необхідно додатково розв’язувати транспортну задачу – мінімізувати сумарні витрати (наприклад, час або відстань) на перевезення дітей до шкіл. Щоб зменшити кількість варіантів для перебору, можна застосувати такий підхід. Дітей із малих хуторів або віддалених фермерських господарств спочатку підвозять до найближчих більших населених пунктів. Таким чином формуються проміжні центри збору. Далі перевезення до шкіл здійснюється вже з цих укрупнених пунктів. Це дозволяє істотно зменшити кількість потенційних місць для розміщення шкіл і, відповідно, знизити розмірність комбінаторної задачі. Після знаходження оптимального розміщення

шкіл на агрегованому рівні можна уточнити розв’язок, врахувавши всі населені пункти та реальні транспортні зв’язки. Однак задача може бути ще складнішою. У деяких випадках доцільно розміщувати школи не безпосередньо в існуючих населених пунктах, а в довільних точках між ними, щоб мінімізувати сумарні витрати на перевезення. Тоді множина можливих варіантів стає неперервною, а сама задача – нелінійною та з декількома локальними мінімумами. У такому випадку розв’язок дискретної (цілочисельної) задачі можна використати як початкове наближення. Надалі, виходячи з цього наближення, здійснюється оптимізація розміщення шкіл у неперервному просторі.

Приклад: Апроксимація багатовимірного розподілу дискретним розподілом меншої розмірності. Розглянемо задачу апроксимації деякого багатовимірного імовірнісного розподілу. Цей розподіл може бути як неперервним, так і дискретним. У випадку дискретного розподілу припустимо, що він містить дуже велику кількість атомів (точок спостережень). Наприклад, це може бути емпіричний розподіл, отриманий на основі вибірки з деякого неперервного розподілу. Потрібно апроксимувати цей розподіл іншим – також дискретним, але з істотно меншою кількістю атомів. При цьому, якщо в емпіричному розподілі всі точки мають однакові ваги (ймовірності), то в апроксимованому розподілі ваги атомів можуть бути різними. Таким чином, задача полягає в оптимальному наближенні вихідного розподілу дискретним розподілом меншої кількості елементів. По суті, це задача кластеризації точок у векторному просторі. Як критерій якості апроксимації має сенс використовувати транспортну відстань, зокрема відстань Канторовича-Рубінштейна (або відстань Васерштейна).

Якщо точки належать деякому метричному простору, то атоми апроксимуючого розподілу можна обмежити точками вихідного розподілу. У цьому випадку задача зводиться до вибору підмножини точок, у яких будуть розміщені атоми нового розподілу, а також до визначення їхніх ваг. Для зниження обчислювальної складності можна застосувати підхід, аналогічний попередньому прикладу. Зокрема:

1. Об’єднати близькі атоми (точки спостережень);
2. Перенести масу малих атомів до найближчих більших або репрезентативних точок;

3. Сформувати укрупнений набір точок із значно меншою кількістю елементів.

   Після цього задача оптимального розміщення атомів апроксимуючого розподілу розв’язується вже на зменшеній множині точок.

Якщо розподіл заданий у лінійному векторному просторі, то атоми апроксимуючого розподілу можуть розташовуватися в довільних точках простору, а не лише серед початкових спостережень. У такому випадку задача стає неперервною, нелінійною та потенційно з декількома мінімумами. Для її розв’язання доцільно:

1. Використати дискретну модель як початкове наближення;
2. Далі застосувати методи неперервної оптимізації для уточнення розміщення атомів і їхніх ваг;

   Якщо початковий розподіл є неперервним, можна діяти поетапно:

1. Згенерувати емпіричний розподіл (вибірку) з деякою кількістю точок;
2. Апроксимувати цей емпіричний розподіл дискретним розподілом із меншою кількістю атомів і нерівними вагами;
3. У процесі надходження нових даних адаптувати апроксимуючий розподіл, уточнюючи як розміщення атомів, так і їхні ваги.

Такий підхід дозволяє побудувати адаптивну модель, що поступово покращує якість апроксимації при надходженні нової інформації.

Приклад: кластеризації векторних даних методом SQC. Одним із базових методів кластеризації є алгоритм K-середніх, який реалізує ітераційний процес поперемінного призначення точок до кластерів та оновлення центроїдів. Проте цей метод має низку недоліків, зокрема високу обчислювальну складність та непридатність до обробки потокових даних без повної повторної кластеризації. У роботі запропоновано метод стохастичної кластеризації, орієнтований на обробку потокових даних. На відміну від класичних підходів, оновлення центроїдів здійснюється ітеративно: при надходженні нового спостереження визначається найближчий центроїд, який коригується у його напрямку. Це дозволяє уникнути повного перегляду вибірки на кожній ітерації, переглядаються тільки точки, прив'язані до цього центроїда. Окрема увага приділяється задачі ініціалізації центроїдів. Запропоновано підхід, що базується на попередній агрегації близьких точок із формуванням зменшеної множини представників. Подальший вибір

початкових центроїдів здійснюється за допомогою методів дискретної оптимізації. Як функціонал якості використовується транспортна відстань (відстань Канторовича-Рубінштейна, або Васерштейна), що дозволяє інтерпретувати задачу кластеризації як задачу оптимального транспорту. Після визначення початкового наближення центроїдів задача уточнення їх розміщення формулюється як задача стохастичного програмування з негладкою неопуклою цільовою функцією. Для її розв'язання застосовуються методи стохастичної оптимізації, зокрема Adam, Adagrad, а також прискорені градієнтні методи. З огляду на негладкість функціонала, у роботі запропоновано процедуру його згладжування. Оцінки градієнта обчислюються за допомогою скінченно-різницевих стохастичних апроксимацій, математичне сподівання яких збігається з градієнтом згладженої функції. Запропонований підхід дозволяє ефективно поєднати кластеризацію, оптимальний транспорт і стохастичну оптимізацію в умовах великорозмірних і змінних даних.

1.2 Про математичні моделі кластеризації даних за принципом центроїдів

Припустимо, що дані представлені елементами в деякому просторі. При кластеризації даних за принципом центроїдів ці елементи групуються за критерієм близькості до деяких точок, які називаються центроїдами. Проблема полягає в оптимальному обрані центроїдів. Коли дані відображаються в деякий метричний простір, в якому не визначені лінійні операції, то постає питання, які точки можна брати в якості центроїдів даних. У цьому випадку центроїди приходиться обирати серед самих даних. Задача оптимальної кластеризації N точок $\{\xi_i, i = 1, ..., N\}$ в m центрів (точок) із того ж набору $\{\xi_i\}$ у метричному просторі з відстанню $\mathrm{dist}(\cdot, \cdot)$ може формулюватися як наступна ЗЦЛП:

$$\sum_{ij} \mathrm{dist}(\xi_i, \xi_j) x_{ij} \to \min_{x_{ij} \geq 0, y_j} \tag{1.1}$$

за обмежень

$$\sum_j x_{ij} = 1/N,\ \sum_i x_{ij} \leq y_j,\ \sum_j y_j \leq m,\ y_j \in \{0, 1\},\ i = 1, ..., N,\ j = 1, ..., N \tag{1.2}$$

Ця задача має комбінаторний характер і вимагає перебору варіантів розміщення m точок по N позиціям $\{\xi_i, i = 1, ..., N\}$ та розв'язання транспортної задачі для кожного такого варіанту. Число таких варіантів дорівнює $N(N-1) \times ... \times (N-m+1)$ та має порядок $N^m = e^{m \log N}$. Якщо дані відображаються (кодуються) у лінійний простір $\mathbb{R}^d$, то в якості центроїдів даних можна обирати проміжні точки $y_j \in \mathbb{R}^d$, але тоді число варіантів вибору центроїдів стає нескінченим. Відповідна задача кластеризації за принципом пошуку центроїдів даних має наступний вигляд:

$$\sum_{ij} \mathrm{dist}(\xi_i, y_j) x_{ij} \to \min_{x_{ij} \geq 0, y_j \in \mathbb{R}^d, q_j \geq 0} \tag{1.3}$$

за обмежень

$$\sum_j x_{ij} = 1/N, \ \sum_i x_{ij} \leq q_j, \ \sum_j q_j = 1 \ i = 1, ..., N, \ j = 1, ..., m. \tag{1.4}$$

Це вже неперервна, нелінійна та неопукла задача оптимізації за лінійних обмежень. При великому числі N елементів даних числове розв'язання цих оптимізаційних задач представляє проблему для сучасних методів та програмних засобів детермінованої оптимізації.

1.3 Постановка задачі машинного навчання «без учителя»

Навчання «без учителя» вирішує задачу виявлення структури в нерозмічених даних, отриманих з невідомого базового розподілу. Формально, маючи набір даних спостережень $X = \{x_1, x_2, \dots, x_N\}$, де $x_i \in \mathbb{R}^d$, метою є вивчення моделі або представлення, яке оптимально пояснює притаманну структуру даних без доступу до явних очікуваних вихідних сигналів [117]. На відміну від парадигм навчання з учителем, які спираються на марковані навчальні приклади, методи «без учителя» повинні виводити закономірності, зв'язки та організаційні принципи безпосередньо з самих даних [53]. Методи навчання «без учителя» покладаються переважно на використання метрик близькості та подібності для виявлення закономірностей і визначення структурної організації даних [117].

Навчання «без учителя» охоплює різноманітне сімейство алгоритмів, спрямованих на виявлення структури латентних ознак з нерозмічених

спостережень через оптимізацію певної метрики. Систематизація даного класу алгоритмів представлена на діаграмі на рисунку 1.1.

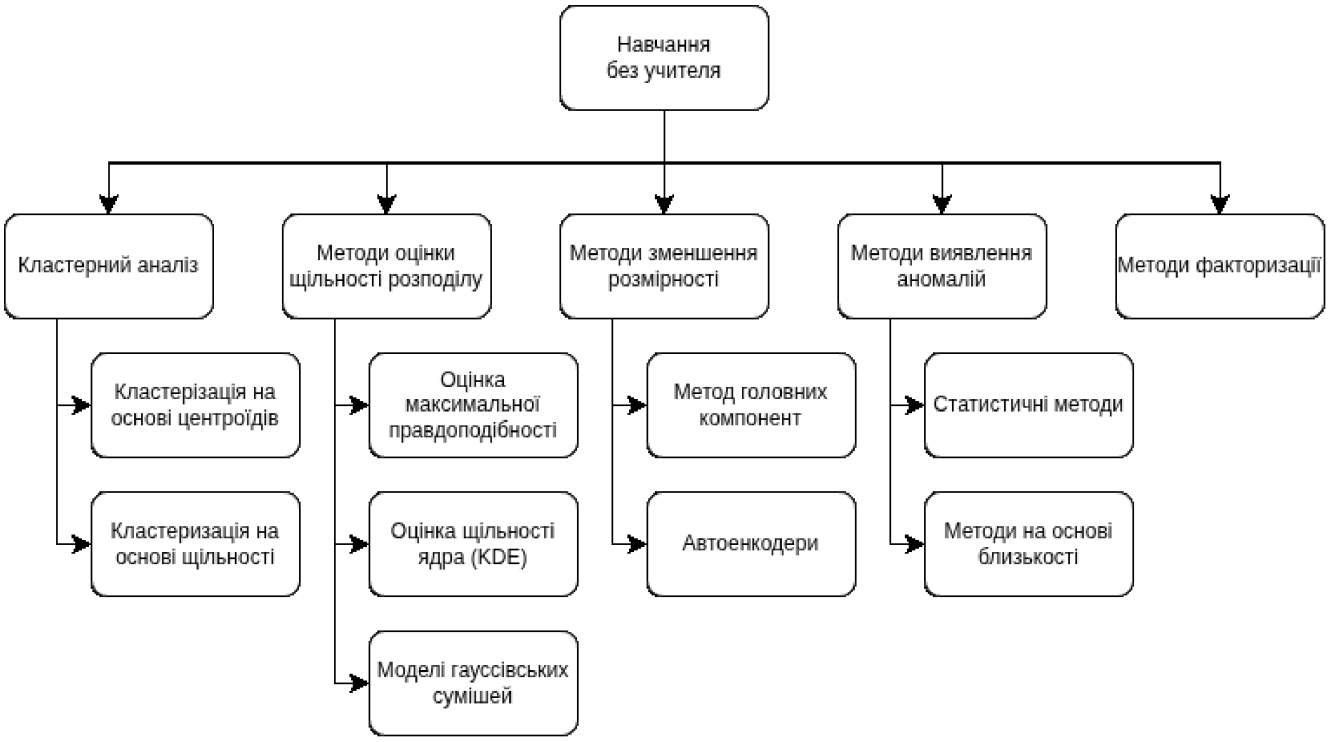

Рис 1.1: Класифікація найпоширеніших методів машинного навчання «без учителя».

Задача зниження розмірності полягає у визначенні низькорозмірних представлень, які зберігають суттєві характеристики даних, зазвичай сформована як пошук матриці проекції $W \in \mathbb{R}^{d \times m}$ з $m \ll d$, яка мінімізує помилку реконструкції $\sum_{i=1}^{N} \|x_i - WW^\top x_i\|_2^2$ за умови ортонормальності $W^\top W = I_m$, як це продемонстровано у методі головних компонент [24]. Більш складні підходи до навчання на многовидах замінюють евклідову відстань геодезичними відстанями або графовими мірами подібності, оптимізуючи цільові функції вигляду $\sum_{i,j} S_{ij} \|f(x_i) - f(x_j)\|^2$, де S_{ij} кількісно характеризує локальну структуру [117].

Задачі оцінювання густини розподілу ймовірності розглядається як знаходження параметрів θ, які максимізують функцію правдоподібності $\sum_{i=1}^{N} \log p_\theta(x_i)$, або за наявності латентних змінних z максимізуючи наступну функцію [114]:

$$\sum_{i=1}^{N} \log \int p_\theta(x_i \mid z) p(z)\, dz \tag{1.5}$$

Матрична факторизація розкладає матрицю даних $X \in \mathbb{R}^{N \times d}$ на фактори нижчого рангу, знаходячи оптимальне значення функції:

$$\min_{W,H} \|X - WH\|_F^2 + \lambda R(W, H) \tag{1.6}$$

де R забезпечує структурні обмеження, такі як невід'ємність або розрідженість [24]. Нарешті, методи виявлення аномалій ідентифікують спостереження, які демонструють низьку міру густини апроксимованого розподілу ймовірності, зазвичай через встановлення порогу $p_\theta(x_i)$ або оцінювання помилки реконструкції відносно нормальних прикладів. Кожен з цих класів задач має спільну математичну основу оптимізації цільової функції, яка вимірює, наскільки добре параметрична модель або представлення відображає базову структуру, присутню в нерозмічених даних, відрізняючись переважно формою цільової функції та накладених обмежень [117].

Одним з яскравих прикладів застосування методів навчання «без учителя» є розпізнавання закономірностей у генетичних послідовностях, де відсутність апріорних знань щодо груп генів вимагає застосування методів розбиття даних на групи [53]. Дані експресії генів зазвичай представлені у вигляді матриць з дійсними значеннями, де рядки-об'єкти відповідають вимірюванням експресії генів у кількох експериментах, а стовпці представляють патерни експресії всіх генів для окремих експериментів з мікрочіпами. Основною метою навчання «без учителя» в цій області є оцінка невідомої функції відображення, яка призначає кластерні мітки генам, що демонструють подібні профілі експресії, тим самим ідентифікуючи корегульовані гени, які можуть брати участь у споріднених біологічних процесах.

Отже, математична постановка задачі машинного навчання «без учителя» включає моделі, які використовують метрики близькості та подібності для виявлення закономірностей на нерозмічених даних. Даний клас методів є важливим для використання в задачах, де марковані навчальні приклади є апріорі неможливими або дуже дорогі для отримання. Така властивість методів дозволяє знаходити розв'язок як для задач відносно низькою розмірністю вхідних даних як генетичний аналіз [53] та сегментації ринку [120], так і для задач з високою розмірністю як аналіз медіаданих [49].

1.4 Існуючі методи кластеризації

Кластерний аналіз являє собою задачу в машинному навчанні «без учителя», метою якого є розділення колекції точок даних на групи на основі мір внутрішньої подібності без опори на заздалегідь визначені мітки класів [129]. Розглянемо набір даних, що складається з N спостережень, позначених як $\Xi = \{\xi_1, \xi_2, \cdots, \xi_N\}$, де кожне $\xi_i \in \mathbb{R}^d$ представляє d-вимірний вектор ознак. Завдання полягає у визначенні розбиття Ξ на K непересічних підмножин $\mathcal{C} = \{C_1, C_2, \cdots, C_K\}$ таких, що $\bigcup_{j=1}^{K} C_j = \Xi$ та $C_i \cap C_j = \emptyset$ для всіх $i \neq j$, де кожна підмножина C_j представляє кластер. Розбиття повинно задовольняти подвійний критерій: спостереження в межах одного кластера демонструють високу подібність, тоді як спостереження в різних кластерах демонструють суттєву відмінність [117].

Метою задачі векторної квантизації, яку також називають кластеризацією на основі центроїдів, є представлення набору даних через скінченну множину елементів-центроїдів, які мінімізують сукупну відстань між точками даних та їхніми найближчими центроїдами [35,70,71]. Формально, враховуючи набір нерозмічених даних Ξ, ми прагнемо визначити скінченну множину елементів $\{y_1, y_2, \cdots, y_K\}$, де $y_j \in \mathbb{R}^d$, яка цільову функцію сумарної квадратичної помилки між кластерами, яка визначена у наступному вигляді:

$$F(\mathcal{C}, \{y_j\}_{j=1}^{K}) = \sum_{j=1}^{K} \sum_{\xi_i \in C_j} \|\xi_i - y_j\|^2 \tag{1.7}$$

де $\|\cdot\|$ позначає евклідову норму. Це формулювання прагне мінімізувати загальну дисперсію всередині кластера, тим самим сприяючи утворенню компактних, добре розділених кластерів. Задача векторної квантизації допускає геометричну інтерпретацію через розбиття простору ознак за методом Вороного [110]. Зокрема, будь-яке розміщення центроїдів $\{y_1, \cdots, y_K\}$ призводить до розбиттів Вороного на $\mathbb{R}^d$, де j-те розбиття визначається як $V_j = \{x \in \mathbb{R}^d,\ \forall k \neq j : \|x - y_j\| \leq \|x - y_k\|\}$. Кожне розбиття Вороного містить усі точки в просторі ознак, які ближче до центроїда y_j, ніж до будь-якого іншого центроїда, тим самим встановлюючи відповідність між центроїдами та кластерами. Ілюстративний приклад розбиття простору ознак за методом Вороного зображений на рисунку 1.2.

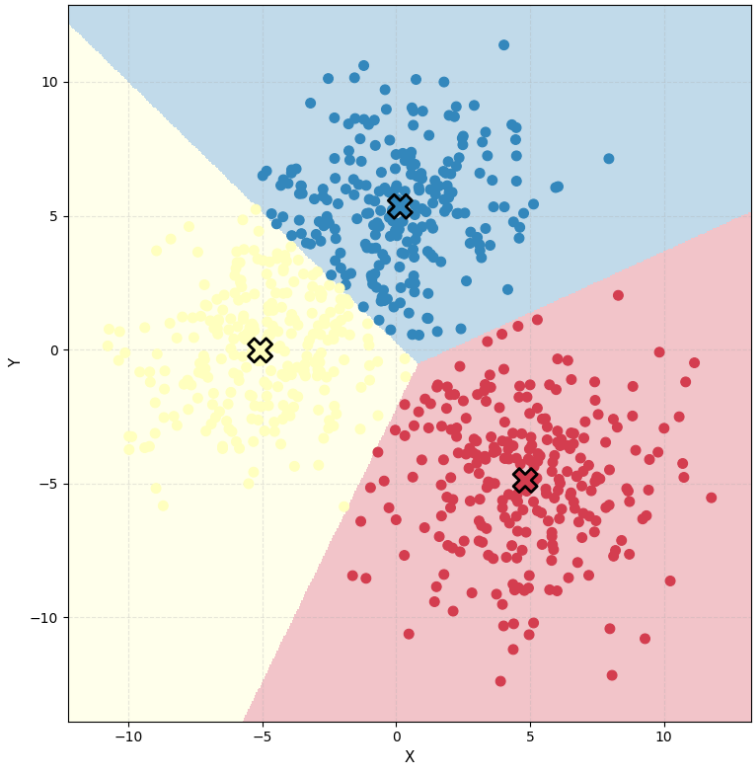

Рис 1.2: Розбиття простору ознак для трьох центроїдів методом векторної квантизації.

Запропонований Стюартом Ллойдом [70], метод K-середніх являє собою евристичний метод для розв'язання задачі векторної квантизації і працює за допомогою стратегії змінної оптимізації, яка ітеративно уточнює як розбиття, так і центроїди [110]. Хоча метод було вперше запропоновано понад п'ять десятиліть тому, він залишається одним з найпоширеніших методів кластеризації завдяки своїй простоті та обчислювальній ефективності [70,71]. З останніми досягненнями в DL, алгоритм продовжує розвиватися та знаходити нові застосування, такі як сприяння індексації векторного простору для ефективного пошуку подібності у високовимірних даних, включаючи пошук зображень [49], незважаючи на початкову вузьку сферу застосування в задачі мінімізації методом найменших квадратів в широтно-імпульсній модуляції [70]. Метод чергується між двома кроками: призначенням спостережень їхнім найближчим центроїдам та перерахуванням центроїдів як середнього значення призначених їм спостережень. Алгоритм працює наступним чином:

1. Ініціалізація: Визначається початкова оцінка K центроїдів $\{y_j^{(0)}\}_{j=1}^K$, зазвичай шляхом випадкової вибірки K елементів з набору даних Ξ або шляхом використання інших методів ініціалізації;
2. Крок присвоєння: Для кожного елементу множини ξ_i визначається приналежність до кластера, ідентифікуючи найближчий центроїд: обчислюється $j^* = \arg\min_{j=1,\dots,K} \|\xi_i - y_j^{(t)}\|^2$ та назначається ξ_i кластеру $C_{j^*}^{(t)}$. Це розбиття відповідає діаграмі Вороного, побудованої поточними центроїдами;
3. Крок оновлення: Кожен центроїд $y_j^{(t+1)}$ переобчислюється як середнє арифметичне всіх спостережень, призначених кластеру $C_j^{(t)}$:

$$y_j^{(t+1)} = \frac{1}{|C_j^{(t)}|} \sum_{\mathbf{x}_i \in C_j^{(t)}} \xi_i \tag{1.8}$$

де $|C_j^{(t)}|$ позначає кількість елементів кластера $C_j^{(t)}$;
4. Ітерація продовжується доки не буде досягнуто критерію збіжності: коли жодна з наступних ітерацій не змінює належність до кластера або коли відносне зменшення цільової функції падає нижче заданого порогу.

Властивістю алгоритму K-середніх є монотонне зменшення цільової функції на кожній ітерації. Формально, нехай $F^{(t)} = F(\mathcal{C}^{(t)}, \{y_j^{(t)}\}_{j=1}^K)$ позначає значення цільової функції (1.7) на ітерації t. Тоді для будь-якої ітерації $t \geq 0$ виконується нерівність $F^{(t+1)} \leq F^{(t)}$ [110]. Ця властивість безпосередньо випливає з будови алгоритму: на кроці присвоєння кожен елемент ξ_i призначається кластеру з найближчим центроїдом, що мінімізує його внесок у цільову функцію при фіксованих центроїдах, а на кроці оновлення кожен центроїд $y_j^{(t+1)}$ обчислюється як середнє арифметичне (1.8), що є оптимальним розв'язком задачі мінімізації суми квадратів відстаней від елементів кластера $C_j^{(t)}$ до центроїда при фіксованому розбитті. Більш детально, для кроку присвоєння маємо:

$$F(\mathcal{C}^{(t)}, \{y_j^{(t)}\}_{j=1}^K) = \sum_{j=1}^K \sum_{\xi_i \in C_j^{(t)}} \|\xi_i - y_j^{(t)}\|^2 \geq \sum_{j=1}^K \sum_{\xi_i \in C_j^{(t+1)}} \|\xi_i - y_j^{(t)}\|^2 \tag{1.9}$$

оскільки кожен елемент призначається кластеру з мінімальною відстанню до центроїда. Аналогічно, для кроку оновлення центроїдів:

$$\sum_{j=1}^{K}\sum_{\xi_i\in C_j^{(t+1)}}\|\xi_i-y_j^{(t)}\|^2\geq\sum_{j=1}^{K}\sum_{\xi_i\in C_j^{(t+1)}}\|\xi_i-y_j^{(t+1)}\|^2=F(\mathcal{C}^{(t+1)},\{y_j^{(t+1)}\}_{j=1}^K) \tag{1.10}$$

що випливає з оптимальності середнього арифметичного як мінімуму суми квадратів відхилень. Комбінуючи обидві нерівності, отримуємо $F^{(t+1)}\leq F^{(t)}$ для всіх $t\geq 0$. Незважаючи на гарантовану монотонну збіжність, алгоритм K-середніх не забезпечує знаходження глобального мінімуму цільової функції (1.7). Ця обмеженість зумовлена некомбінаторною властивістю задачі векторної квантизації: цільова функція $F(\mathcal{C},\{y_j\}_{j=1}^K)$ є неопуклою функцією параметрів розбиття та центроїдів [117]. Формально, для двох різних положень центроїдів $\{y_j\}_{j=1}^K$ та $\{y'_j\}_{j=1}^K$ та відповідних їм розбиттів $\mathcal{C}$ та $\mathcal{C}'$, не обов'язково виконується властивість опуклості:

$$\begin{aligned}F(\lambda\mathcal{C}+(1-\lambda)\mathcal{C}',\{\lambda y_j+(1-\lambda)y'_j\}_{j=1}^K)\not\leq\\ \not\leq\lambda F(\mathcal{C},\{y_j\}_{j=1}^K)+(1-\lambda)F(\mathcal{C}',\{y'_j\}_{j=1}^K)\end{aligned} \tag{1.11}$$

для $\lambda\in(0,1)$. Більше того, задача знаходження глобального оптимуму розбиття N елементів на K кластерів є NP-складною [117], оскільки кількість можливих розбиттів зростає експоненційно з розміром набору даних. Внаслідок неопуклості цільової функції, алгоритм K-середніх гарантує збіжність лише до локального мінімуму, який суттєво залежить від вибору початкового наближення центроїдів $\{y_j^{(0)}\}_{j=1}^K$ [70,117]. Ця чутливість до ініціалізації проявляється у тому, що різні початкові наближення можуть призводити до якісно різних локальних мінімумів з істотно відмінними значеннями цільової функції.

Метод K-середніх окрім властивості монотонної збіжності (1.10) також гарантує збіжність за скінченну кількість ітерацій, що виводиться зі самої постановки задачі: кількість можливих розбиттів множини з N елементів на K непорожніх непересічних підмножин визначається числом Стірлінга другого роду [110]:

$$S(N,K)=\frac{1}{K!}\sum_{j=0}^{K}(-1)^{K-j}\binom{K}{j}j^N \tag{1.12}$$

яке хоча й є великим, але залишається скінченним для будь-яких скінченних множин елементів $\{\xi_i\}_{i=1}^N$ та $\{y_j\}_{j=1}^K$. Оскільки на кожній ітерації алгоритм K-середніх визначає конкретне розбиття множини спостережень Ξ та відповідні центроїди, а

цільова функція монотонно не зростає і є обмеженою знизу нулем, алгоритм не може нескінченно переходити між різними наближеннями. Формально, якщо позначити через $(\mathcal{C}^{(t)}, \{y_j^{(t)}\}_{j=1}^K)$ наближення на ітерації t, то існує скінченне число $T < \infty$ таке, що $\mathcal{C}^{(T)} = \mathcal{C}^{(T+1)}$ та $\{y_j^{(T)}\}_{j=1}^K = \{y_j^{(T+1)}\}_{j=1}^K$, оскільки монотонне зменшення дискретного скінченного набору значень $\{F^{(0)}, F^{(1)}, F^{(2)}, ...\}$ неминуче досягає стаціонарної точки [110]. Таким чином, алгоритм завершує роботу, коли жодна зміна розбиття не призводить до зменшення цільової функції, що відповідає умові локального оптимуму. На практиці збіжність часто досягається значно швидше, ніж могла б передбачити верхня межа $S(N, K)$, особливо для добре розділених кластерів, проте теоретична гарантія скінченного часу збіжності залишається властивістю методу [70].

На практиці класичний метод K-середніх має певний ряд обмежень, як залежність від Евклідової відстані [25,46] та чутливість до статистичного шуму [21], що мотивували розробку численних модифікацій, які використовують альтернативні метричні простори та стратегії оптимізації. Структура алгоритму, що чергує кроки присвоєння та оновлення центроїдів, залишається незмінною у цих варіантах, проте вибір міри відстані та методу обчислення центроїдів суттєво впливає на властивості отриманих розбиттів. Міри відстані, відмінні від квадрату евклідової норми, можуть бути використані в рамках загальної парадигми K-середніх, що призводить до алгоритмів з якісно різними характеристиками збіжності та стійкості.

Метод K-медіан представляє собою узагальнення класичного алгоритму K-середніх, де замість квадрату евклідової відстані використовується L_1-норма, а центроїди кластерів визначаються як покомпонентні медіани замість середніх арифметичних [103]. Формально, цільова функція методу K-медіан задається у вигляді:

$$F_{\text{med}}(\mathcal{C}, \{y_j\}_{j=1}^K) = \sum_{j=1}^{K} \sum_{\xi_i \in C_j} \|\xi_i - y_j\|_1 = \sum_{j=1}^{K} \sum_{\xi_i \in C_j} \sum_{l=1}^{d} |\xi_i^{(l)} - y_j^{(l)}| \tag{1.13}$$

де $\|\cdot\|_1$ позначає L_1-норму, $\xi_i^{(l)}$ та $y_j^{(l)}$ позначають l-ту компоненту відповідно елемента ξ_i та центроїда y_j. Оптимальний центроїд кластера C_j у методі K-медіан

визначається як покомпонентна медіана всіх спостережень, призначених цьому кластеру [103,117]:

$$y_j^{(l)} = \mathrm{med}\{\xi_i^{(l)} : \xi_i \in C_j\}, \quad l = 1, \dots, d \tag{1.14}$$

де $\mathrm{med}(\cdot)$ позначає медіану множини значень. Ця властивість випливає з відомого факту, що медіана мінімізує суму абсолютних відхилень: для будь-якої скінченної множини дійсних чисел $\{a_1, \dots, a_m\}$ значення $\mathrm{med}\{a_1, \dots, a_m\}$ є розв'язком задачі $\arg\min_{c \in \mathbb{R}} \sum_{i=1}^{m} |a_i - c|$ [103,117]. Алгоритм K-медіан зберігає структуру ітеративного уточнення класичного методу K-середніх, чергуючи крок присвоєння, де кожен елемент ξ_i призначається кластеру з найближчим центроїдом відносно L_1-норми, та крок оновлення, де центроїди перераховуються згідно з (1.14).

Перевагою методу K-медіан над класичним методом K-середніх є підвищена стійкість до статистичного шуму у даних [62]. Ця властивість має математичне обґрунтування у теорії статистики, оскільки медіана є точкою розбиття $1/2$, тоді як середнє арифметичне є точку розбиття $1/N$, що прямує до нуля при зростанні розміру вибірки [103,117]. Формально, вплив викиду на центроїд у методі K-середніх є необмеженим: якщо елемент $\xi_i \in C_j$ замінити на $\xi_i + \delta e$ для деякого одиничного вектора e та необмежено зростаючого $\delta \to \infty$, то відповідний центроїд кластера зміщується пропорційно $\delta/|C_j|$, тобто$y_j \to y_j + (\delta/|C_j|)e$, що призводить до необмеженого зростання цільової функції. Навпаки, у методі K-медіан такий статистичний шум впливає лише на порядкову статистику елементів кластера, залишаючи медіану незмінною доти, доки шум не перевищує половину елементів кластера. Проте, як було зазначено в дослідженні [36], метод K-медіан демонструє стійкість до статистичного шуму лише у спеціальних випадках, а в загальному випадку цільова функція (1.13) може бути настільки ж чутлива до шуму, як і у класичному методі K-середніх [117]. Це пояснюється тим, що хоча медіана є стійкою мірою центральної тенденції для одновимірних даних, спільний вибір медіан для багатовимірних даних не обов'язково зберігає стійкі властивості, оскільки шум може істотно вплинути на призначення інших елементів до кластерів через зміну розбиттів Вороного.

Альтернативним підходом до модифікації класичного методу K-середніх є метод гармонійних K-середніх, запропонований для підвищення стійкості алгоритму до вибору початкових центроїдів [74]. Цей метод базується на концепції

гармонічного середнього та замінює традиційну цільову функцію на вираз, що враховує відстані від кожного елемента до всіх центроїдів одночасно. Для скалярної множини $\Xi = \{\xi_1, \cdots, \xi_N\}$ гармонічне середнє визначається як:

$$\bar{\xi}_H = \frac{N}{\sum_{i=1}^{N} \xi_i^{-1}} \tag{1.15}$$

що є оберненою величиною середнього арифметичного обернених значень. У контексті кластеризації цей принцип узагальнюється на багатовимірний випадок через зважене гармонічне середнє квадратів відстаней від елементів до центроїдів. Цільова функція методу гармонійних K-середніх формулюється у наступному вигляді [74]:

$$F_{\text{HKM}}(\mathcal{C}, \{y_j\}_{j=1}^{K}) = \sum_{i=1}^{N} \frac{K}{\sum_{j=1}^{K} \frac{1}{\|\xi_i - y_j\|^2}} = \sum_{i=1}^{N} \frac{K}{\sum_{j=1}^{K} \frac{1}{(\xi_i - y_j)^\top (\xi_i - y_j)}} \tag{1.16}$$

Суттєвою відмінністю від класичного методу K-середніх є те, що у виразі (1.16) для кожного елемента ξ_i враховуються відстані до всіх K центроїдів одночасно, а не лише до найближчого центроїда, що забезпечує більш глобальне уявлення про розташування елемента відносно всіх кластерів. Оптимізація цільової функції (1.16) методом часткових похідних призводить до наступної оцінки центроїдів [74]:

$$y_j = \frac{\sum_{i=1}^{N} \|\xi_i - y_j\|^{-4} \left(\sum_{s=1}^{K} \|\xi_i - y_s\|^{-2} \right)^{-2} \xi_i}{\sum_{i=1}^{N} \|\xi_i - y_j\|^{-4} \left(\sum_{s=1}^{K} \|\xi_i - y_s\|^{-2} \right)^{-2}} \tag{1.17}$$

що у матричній формі для Фробеніусової норми набуває вигляду, де $\mathrm{Sp}(\cdot)$ позначає слід матриці. Рівень нечіткої належності елемента ξ_i до кластера j на основі оцінки (1.17) визначається як:

$$\mu_j(\xi_i) = \frac{\|\xi_i - y_j\|^{-4}}{\sum_{s=1}^{K} \|\xi_i - y_s\|^{-4}} = \frac{\mathrm{Sp}((\xi_i - y_j)(\xi_i - y_j)^\top)^{-2}}{\sum_{s=1}^{K} \mathrm{Sp}((\xi_i - y_s)(\xi_i - y_s)^\top)^{-2}} \tag{1.18}$$

що має структуру, подібну до оцінки належності у методі нечітких K-середніх.

Метод гармонійних K-середніх демонструє значно меншу чутливість до вибору початкових центроїдів порівняно з класичним алгоритмом K-середніх, що є особливо важливим при аналізі великих масивів даних невідомої структури [40,132]. Ця властивість зумовлена тим, що цільова функція (1.16) враховує відстані до всіх центроїдів одночасно, надаючи більшу вагу елементам, що знаходяться на великих відстанях від поточного розміщення центроїдів через

обернено-квадратичну залежність у знаменнику. Формально, якщо елемент ξ_i знаходиться далеко від усіх поточних центроїдів, тобто $\min\limits_j \|\xi_i - y_j\| \gg 0$, то його внесок у цільову функцію $F_{\text{НКМ}}$ зростає пропорційно $\sum_{j=1}^{K} \|\xi_i - y_j\|^{-2}$, що стимулює переміщення центроїдів у напрямку таких віддалених елементів. Показано, що метод гармонійних K-середніх призводить до результатів, близьких до тих, що забезпечує класичний метод K-середніх, однак процедура є значно менш чутливою до початкової специфікації прототипів-центроїдів, що є особливо зручним при обробці довільних масивів даних [74]. З обчислювальної точки зору, процедура гармонічних K-середніх є неістотно складнішою за стандартний метод K-середніх, оскільки обчислення сліду матриці не вимагає значних обчислювальних ресурсів, є лінійною функцією та відповідно не збільшує складність алгоритму.

Концептуально відмінним підходом до векторної квантизації є метод нечітких K-середніх, який замість жорсткого призначення кожного елемента до єдиного кластера допускає часткову належність елементів до декількох кластерів одночасно [12,29,110]. У методі нечітких K-середніх кожному елементу ξ_i ставиться у відповідність вектор належностей $\mathbf{u}_i = (u_{i1}, \ldots, u_{iK})^\top$ представляє ступінь належності елемента ξ_i до кластера j, що задовольняє умовам нормалізації:

$$\sum_{j=1}^{K} u_{ij} = 1, \quad i = 1, \ldots, N, \qquad \sum_{i=1}^{N} u_{ij} > 0, \quad j = 1, \ldots, K \tag{1.19}$$

де перша умова забезпечує повне розподілення належності кожного елемента між усіма кластерами, а друга гарантує непорожність кожного кластера. Множина всіх таких векторів належностей утворює матрицю належності $U \in [0,1]^{N \times K}$, що представляє нечітке розбиття набору даних. Функція належності u_{ij} інтерпретується як ймовірність, що елемент ξ_i належить кластеру j, причому сума за всіма кластерами дорівнює одиниці [29,110].

Цільова функція методу нечітких K-середніх визначається як зважена сума квадратів відстаней між елементами та центроїдами, де ваги задаються степенями функцій належності [12,29]:

$$F_{\text{FCM}}(\{y_j\}_{j=1}^{K}, U) = \sum_{i=1}^{N}\sum_{j=1}^{K} u_{ij}^q \|\xi_i - y_j\|^2 = \sum_{i=1}^{N}\sum_{j=1}^{K} u_{ij}^q (\xi_i - y_j)^\top (\xi_i - y_j) \tag{1.20}$$

де параметр $q > 1$ називається коефіцієнтом нечіткості та контролює ступінь розмитості розбиття: при $q \to 1$ нечітке розбиття прямує до жорсткого, тоді як при великих значеннях q належності елементів наближаються до рівномірного розподілу $u_{ij} \approx 1/K$ для всіх i, j [110]. Оптимізація цільової функції (1.20) за умов (1.19) виконується методом множників Лагранжа. Визначається функція Лагранжа:

$$\mathcal{L}(\{y_j\}_{j=1}^{K}, U, \{\lambda_i\}_{i=1}^{N}) = \sum_{i=1}^{N}\sum_{j=1}^{K} u_{ij}^{q}\|\xi_i - y_j\|^2 - \sum_{i=1}^{N}\lambda_i\left(\sum_{j=1}^{K} u_{ij} - 1\right) \tag{1.21}$$

де λ_i є множниками Лагранжа, що забезпечують виконання обмежень нормалізації. Прирівнюючи частинну похідну $\partial\mathcal{L}/\partial u_{rs} = 0$ для довільних r, s та припускаючи $\xi_r \neq y_s$, отримуємо $qu_{rs}^{q-1}\|\xi_r - y_s\|^2 - \lambda_r = 0$, звідки [110]:

$$u_{rs} = \left(\frac{\lambda_r}{q\|\xi_r - y_s\|^2}\right)^{\frac{1}{q-1}} \tag{1.22}$$

Підставляючи цей вираз у перше обмеження (1.19), визначаємо множник λ_r та отримуємо оптимальну функцію належності:

$$u_{ij} = \frac{1}{\sum_{s=1}^{K}\left(\frac{\|\xi_i - y_j\|}{\|\xi_i - y_s\|}\right)^{\frac{2}{q-1}}}, \quad \forall j,\ \xi_i \neq y_j \tag{1.23}$$

У випадку співпадіння елемента з одним або декількома центроїдами, тобто коли множина $I_i = \{s : y_s = \xi_i\} \neq \emptyset$, належність визначається як $u_{ij} = 1/|I_i|$ якщо $j \in I_i$ та $u_{ij} = 0$ в іншому випадку. Оптимальні центроїди кластерів визначаються прирівнюванням до нуля частинної похідної функції Лагранжа за центроїдами: $\partial\mathcal{L}/\partial y_j = 2\sum u_{ij}^{q}(y_j - \xi_i) = 0$, що призводить до зваженого середнього арифметичного:

$$y_j = \frac{\sum_{i=1}^{N} u_{ij}^{q}\xi_i}{\sum_{i=1}^{N} u_{ij}^{q}}, \quad j = 1, \ldots, K \tag{1.24}$$

Метод нечітких K-середніх зберігає структуру ітеративної оптимізації класичного алгоритму, чергуючи крок оновлення належностей згідно з формулою (1.23) при фіксованих центроїдах та крок оновлення центроїдів згідно з (1.24) при фіксованих асоціаціях з центроїдами. Подібно до класичного методу K-середніх, метод нечітких K-середніх гарантує монотонне зменшення цільової функції на кожній ітерації та збіжність до локального мінімуму за скінченну кількість кроків [29,110,117]. Проте метод нечітких K-середніх може розглядатися як метод, що не

абсолютно призначає об'єкти до кластера, а визначає деяку ймовірність виникнення об'єктів з різних кластерів. Формально, нечітке розбиття надає більш гнучке представлення структури даних порівняно з жорстким розбиттям, дозволяючи елементам, що знаходяться на межі між кластерами, мати істотну належність до декількох кластерів одночасно. Величина дисперсії j-го кластера з центром y_j у нечіткій кластеризації обчислюється як зважена дисперсія:

$$\sigma_j^2 = \frac{1}{\sum_{i=1}^{N} u_{ij}^q} \sum_{i=1}^{N} u_{ij}^q \|\xi_i - y_j\|^2, \quad j = 1, \dots, K \tag{1.25}$$

що узагальнює класичну внутрішньокластерну дисперсію на випадок нечіткої належності. Вибір оптимального значення параметра нечіткості q, при якому цільова функція (1.20) досягає глобального мінімуму, є складною задачею, і існують численні дослідження цієї проблеми [38,126]. У прикладних дослідженнях найчастіше використовується значення $q \in [1.5, 2.5]$, причому $q = 2$ є найпоширенішим вибором завдяки балансу між гнучкістю нечіткого розбиття та обчислювальною стійкістю [110]. Однак, незважаючи на теоретичні переваги нечіткого підходу, у майже всіх великих застосуваннях нечіткі розбиття не можуть бути ефективно використані, і об'єкт призначається до групи, пов'язаної з найбільшим значенням належності, що фактично перетворює нечітке розбиття на жорстке та нівелює переваги методу [117].

Концептуально близьким до методів векторної квантизації, але побудованим у рамках ймовірнісного моделювання, є метод GMM, який розглядає задачу кластеризації як задачу оцінювання щільності розподілу ймовірності спостережень. На відміну від нечітких K-середніх, де належність елементів до кластерів інтерпретується евристично через ваги u_{ij}, метод GMM припускає, що спостережувані дані породжуються змішаним розподілом K нормальних компонент, кожна з яких описує окрему латентну групу [47]. Формально, щільність розподілу даних у моделі GMM визначається як зважена сума щільностей компонент:

$$p(\xi \mid \theta) = \sum_{j=1}^{K} \pi_j \, \phi(\xi \mid \mu_j, \Sigma_j) \tag{1.26}$$

де $\theta = \{(\pi_j, \mu_j, \Sigma_j)\}_{j=1}^{K}$ позначає вектор параметрів моделі, $\pi_j \geq 0$ – ваги компонент із обмеженням $\sum_{j=1}^{K} \pi_j = 1$ а $\phi(\xi \mid \mu_j, \Sigma_j)$ – щільність багатовимірного нормального розподілу з вектором середніх $\mu_j \in \mathbb{R}^d$ та коваріаційною матрицею $\Sigma_j \in \mathbb{R}^{d \times d}$. Оцінка параметрів моделі (1.26) методом максимальної правдоподібності для незалежної вибірки $\{\xi_i\}_{i=1}^{N}$ зводиться до мінімізації від'ємної логарифмічної функції правдоподібності:

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \log \left(\sum_{j=1}^{K} \pi_j \, \phi(\xi_i \mid \mu_j, \Sigma_j) \right) \to \min_{\theta} \tag{1.27}$$

Стандартним методом розв'язання задачі (1.27) є алгоритм EM, який чергує крок обчислення апостеріорних ймовірностей належності елементів до компонент та крок максимізації нижньої межі логарифмічної правдоподібності за параметрами [47]. Метод K-середніх можна розглядати як спеціальний випадок EM-алгоритму для GMM зі сферичними рівноваговими компонентами та твердим призначенням точок до компонент, що формально відповідає границі $\Sigma_j = \sigma^2 I$ при $\sigma \to 0$. Подібно до методу K-середніх, EM-алгоритм гарантує монотонне неспадання функції правдоподібності на кожній ітерації та збіжність до стаціонарної точки за скінченну кількість кроків, проте не гарантує знаходження глобального максимуму [47]. Перевагою імовірнісного підходу є узагальнена інтерпретація апостеріорних ймовірностей як ступенів належності та можливість моделювання кластерів довільної форми за допомогою повних коваріаційних матриць Σ_j, що дозволяє врахувати анізотропію та корельованість ознак, недоступну в методах векторної квантизації з евклідовою відстанню.

Отже, методи векторної квантизації та споріднені імовірнісні підходи представляють собою клас алгоритмів машинного навчання «без учителя», що забезпечують розбиття даних на групи через ітеративну оптимізацію центроїдів та призначень до кластерів. Класичний метод K-середніх та його численні модифікації, включаючи K-медіани, гармонічних K-середніх, нечіткі K-середніх та GMM, демонструють різні компромісні рішення між обчислювальною складністю, стійкістю до статистичного шуму та чутливістю до ініціалізації, зберігаючи при цьому гарантію монотонної збіжності до локального мінімуму за скінченну кількість ітерацій. Незважаючи на різні характеристики методів K-середніх, варто

зауважити, що дані методи були створені для вирішення задач, де за замовчуванням вхідний набір даних $\{\xi_i\}_{i=1}^N$ залишається фіксованим. В наступному розділі буде розглянуто які існують обмеження у даних методів в умовах потоків даних, коли дані додаються до початкового набору з часом.

1.5 Неопуклість цільової функції методів кластеризації

Як було зазначено у попередньому розділі, всі розглянуті методи кластеризації – класичний K-середніх, K-медіан, гармонічних K-середніх, нечіткі K-середні та GMM – гарантують лише монотонне зменшення відповідної цільової функції та збіжність до локального мінімуму за скінченну кількість ітерацій. Спільною рисою цих алгоритмів є те, що їх якість суттєво залежить від вибору початкового наближення центроїдів, і різні ініціалізації можуть призводити до якісно відмінних розв'язків.

Розглянемо детальніше структуру неопуклості цільової функції методу K-середніх. Цільова функція (1.7) залежить як від дискретного розбиття $\mathcal{C}$, так і від неперервних центроїдів $\{y_j\}_{j=1}^K$, причому при фіксованому розбитті F є опуклою (а саме сумою квадратичних функцій) за центроїдами, а при фіксованих центроїдах оптимальне розбиття визначається однозначно через найближчий центроїд. Однак спільна задача оптимізації за обома множинами змінних не є опуклою [3,117]. Найпростішим проявом неопуклості є інваріантність цільової функції відносно перестановки центроїдів: якщо $\{y_j^*\}_{j=1}^K$ є локальним мінімумом, то будь-яка перестановка $\{y_{\pi(j)}^*\}_{j=1}^K$ для $\pi \in S_K$ дає те саме значення цільової функції, що породжує K еквівалентних локальних мінімумів. Це означає, що для опуклої комбінації двох таких мінімумів $\lambda\{y_j^*\}_{j=1}^K + (1-\lambda)\{y_{\pi(j)}^*\}_{j=1}^K$ значення цільової функції може суттєво перевищувати лінійну інтерполяцію значень у кінцевих точках, що безпосередньо суперечить означенню опуклості [117]. Окрім симетрії перестановки, існують також нееквівалентні локальні мінімуми, що відповідають різним розбиттям набору даних і дають різні значення цільової функції. Формально, для квадратичної помилки векторної квантизації справедлива тотожність [3]:

$$\sum_{j=1}^{K}\sum_{\xi_i\in C_j}\|\xi_i-y_j\|^2=\sum_{j=1}^{K}\frac{1}{2|C_j|}\sum_{\xi_i,\xi_l\in C_j}\|\xi_i-\xi_l\|^2 \tag{1.28}$$

яка показує, що мінімізація суми квадратів відстаней до центроїдів еквівалентна мінімізації суми парних квадратичних відстаней всередині кластерів, нормованих за значенням ступенів. Це формулювання робить задачу векторної квантизації комбінаторною: для N елементів та K кластерів кількість можливих розбиттів задається числом Стірлінга другого роду (1.12), яке зростає експоненційно з N.

Обчислювальна складність задачі знаходження глобального мінімуму цільової функції векторної квантизації була предметом численних досліджень [3,72]. В роботі [3] було доведено, що задача мінімізації суми квадратів евклідових відстаней є NP-складною у просторах загальної розмірності навіть для випадку $K=2$ кластерів, шляхом зведення від задачі найщільнішого розрізу графа. Це означає, що за умови $\mathrm{P}\neq\mathrm{NP}$ не існує детермінованого алгоритму поліноміальної складності, який гарантовано знаходить глобальний оптимум задачі векторної квантизації у просторі $\mathbb{R}^d$ при d, що входить до розміру вхідних даних. Більше того, задача K-середніх залишається NP-складною навіть на площині $\mathbb{R}^2$ [72]. Формально, для заданої скінченної множини точок $\Xi\subset\mathbb{R}^2$ з раціональними координатами, цілого числа $K\geq 1$ та раціональної межі R, задача визначення існування K центроїдів, що забезпечують значення цільової функції (1.7) не більше за R, є NP-складною [72]. Ці результати спільно встановлюють, що задача оптимальної векторної квантизації є обчислювально нерозв'язною за поліноміальний час як у фіксованій малій розмірності з варіативним K, так і у загальній розмірності з фіксованим K. Як наслідок, метод K-середніх та його модифікації, що ґрунтуються на ітеративній локальній оптимізації, можуть забезпечувати лише наближені розв'язки, причому якість таких наближень залежить від стратегії ініціалізації центроїдів [70,117].

Неопуклість цільової функції та її наслідки для збіжності локальних методів оптимізації не обмежуються методом K-середніх та його прямими модифікаціями, а є характерною властивістю більш широкого класу моделей кластеризації. Зокрема, від'ємна логарифмічна функція правдоподібності моделі GMM (1.26), визначена у попередньому розділі, також є неопуклою за параметрами $\theta=\{(\pi_j,\mu_j,\Sigma_j)\}_{j=1}^{K}$, причому неопуклість має складнішу структуру, ніж у методі K-

середніх. Окрім тривіальної інваріантності відносно перестановки компонент, що породжує K еквівалентних локальних мінімумів, для GMM з $M \geq 3$ компонентами існують суттєво нееквівалентні локальні мінімуми, значення цільової функції в яких може бути довільно гіршим за глобальний оптимум. Формально, для будь-якого $M \geq 3$ та будь-якої константи $C_{\text{gap}} > 0$ існує добре розділена рівновагова суміш M сферичних розподілів з одиничною дисперсією $\text{GMM}(\mu^*)$ та локальний максимум μ_{loc} функції правдоподібності, такий що [47]:

$$\mathcal{L}(\mu_{\text{loc}}) \leq \mathcal{L}(\mu^*) - C_{\text{gap}} \tag{1.29}$$

де $\mathcal{L}$ позначає популяційну логарифмічну правдоподібність, а μ^* – істинні центри компонент, що відповідають глобальному максимуму. Геометрично така ситуація виникає, коли при певному наближенні істинних центрів декілька оцінюваних центрів зосереджуються в околі одного істинного кластера, тоді як інший істинний кластер представлений єдиним усередненим центром, що покриває одночасно дві справжні компоненти [47]. Більше того, EM-алгоритм та його градієнтний варіант з рівномірною випадковою ініціалізацією центрів з самої вибірки збігаються до менш оптимального значення з ймовірністю принаймні $1 - e^{-\Omega(M)}$. Це означає, що навіть для добре розділених GMM ймовірність успіху одного запуску EM-алгоритму експоненційно спадає зі зростанням кількості компонент, і для досягнення розв'язку з прийнятною якістю необхідне експоненційне число повторних запусків з різних випадкових ініціалізацій.

Отже, неопуклість цільових функцій методів кластеризації, як для прямої задачі векторної квантизації (1.7), так і для імовірнісної моделі GMM (1.26), є властивістю, що визначає їх обчислювальну поведінку. NP-складність задачі знаходження глобального мінімуму [3,72] та доведене існування довільно поганих локальних оптимумів у моделі GMM [47] разом встановлюють, що жодна модифікація методу K-середніх з покращеним правилом оновлення центроїдів – будь то K-медіани, гармонічних K-середніх чи нечіткі K-середні – не може розв'язати проблему неопуклості. Це мотивує два напрямки подальшого дослідження. По-перше, необхідні розумні стратегії ініціалізації центроїдів, які забезпечують потрапляння алгоритму у басейн притягання якісного локального мінімуму [47,117]. По-друге, для переходу від локальних мінімумів до глобальних доцільно застосовувати методи стохастичної оптимізації, які за рахунок

випадкових збурень градієнта здатні залишати неоптимальні стаціонарні точки. Саме ця властивість стохастичних методів буде використана у запропонованому методі стохастичної кластеризації, де адаптивне оновлення центроїдів за стохастичними оцінками градієнта дозволить одночасно опрацьовувати неопуклість цільової функції та забезпечувати стійкість до дрейфу розподілу даних, що розглядається у наступному розділі.

1.6 Кластеризація в умовах потоків даних

Алгоритми розглянуті в роботі застосовуються до задач детермінованої оптимізації, де цільова функція і (або) обмеження є точно визначеними, а множина вхідний набір даних $\{\xi_i\}_{i=1}^N$ є фіксованим. Проте в задачах, де розподіл даних може змінюватись з часом через сезонні чи соціологічні коливання, необхідно знайти оптимальний розв’язок цільової функції в умовах невизначеності та статистичного шуму, наприклад в задачі аналізу ринку [120]. В цьому контексті задача приймає стохастичний характер, де моделюється оптимізація відносно випадкових величин, і визначається в наступному вигляді [32,33,80,83,113]:

$$\min_{x\in X}[f(x) = \mathbb{E}F(x,\xi) = \int_{\xi\in\Xi} F(x,\xi)P(d\xi), \quad X \subset \mathbb{R}^d \tag{1.30}$$

де $\mathbb{E}$ є оператором математичного сподівання, $f : X \to \mathbb{R}^d$ – цільова функція з областю визначення $X \subset \mathbb{R}^d$. $F : X \times \Xi \to \mathbb{R}^1$ – опукла та диференційовна функція, яка залежить від визначеної змінної $x \in X$ та стохастичної змінної $\xi \in \Xi$, визначеної на просторі (Ξ, Σ, P).

Подальші властивості задачі стохастичної оптимізації (1.30) та методи знаходження розв'язку будуть розглянуті в наступних розділах, але в даному розділі буде зроблено акцент на проблемі розподільного дрейфу яка утворюється в задачі машинного навчання в стохастичних умовах. Передумовою класичних підходів до стохастичної оптимізації є припущення про стаціонарність розподілу ймовірностей P випадкової величини ξ [16]. Формально, це означає, що якщо $\{\xi_1, \xi_2, \dots, \xi_N\}$ представляє послідовність незалежних спостережень, то кожне ξ_i вибирається з єдиного спільного розподілу P, який не змінюється з часом. За такої умови, оптимальне розв'язання $x^* \in X$ задачі (1.30), визначене як

$x^* = \arg\min_{x \in X} \mathbb{E}_{\xi \sim P}[F(x, \xi)]$, залишається незмінним протягом усього процесу оптимізації. Проте у реальних прикладних задачах, зокрема в навчанні DNN та адаптивному аналізу сигналів, це припущення часто порушується: статистичний розподіл даних може еволюціонувати з часом внаслідок сезонних змін, соціологічних тенденцій або технологічних факторів [7,120].

Для формального опису цього явища, введемо поняття часозалежних розподілів. Нехай $\{P_t\}_{t=1}^{\infty}$ позначає послідовність розподілів ймовірностей над простором Ξ, де індекс $t \in \mathbb{N}$ відповідає дискретному часовому моменту. Відповідно, задача стохастичної оптимізації з часозалежними розподілами формулюється як [16]:

$$\min_{x \in X} F_t(x) := \mathbb{E}_{\xi \sim P_t}[F(x, \xi)] = \int_{\xi \in \Xi} F(x, \xi) P_t(d\xi) \tag{1.31}$$

В цьому контексті оптимальне розв'язання також стає часозалежним: $x_t^* = \arg\min_{x \in X} F_t(x)$. Послідовність $\{x_t^*\}_{t=1}^{\infty}$ визначає траєкторію динамічних оптимальних точок, яка відстежує еволюцію розподілу даних. Відмінність між статичною постановкою (1.30) та динамічною постановкою (1.31) полягає у виборі критерію ефективності алгоритму. У статичному випадку критерієм є відхилення від єдиної оптимальної точки x^*, тоді як у динамічному випадку необхідно порівнювати ефективність алгоритму з траєкторією динамічних оптимумів $\{x_t^*\}_{t=1}^{\infty}$ [16]. Для кількісної оцінки тимчасових змін розподілу даних, вводиться поняття дрейфу оптимальних точок. Формально, дрейф на горизонті T визначається як сумарна зміна оптимальних розв'язань між послідовними часовими моментами:

$$\Delta(T) := \sum_{t=2}^{T} \|x_{t-1}^* - x_t^*\|_2 \tag{1.32}$$

де $\|\cdot\|_2$ позначає евклідову норму у $\mathbb{R}^d$. Величина $\Delta(T)$ вимірює кумулятивне переміщення оптимальних точок у просторі параметрів і таким чином характеризує інтенсивність змін у розподілі даних. Якщо $\Delta(T) = o(T)$, тобто дрейф є сублінійним відносно горизонту часу, це означає, що розподіл змінюється відносно повільно. Навпаки, якщо $\Delta(T) = \Theta(T)$ або $\Delta(T) = \omega(T)$, розподіл змінюється з постійною або зростаючою швидкістю, що робить задачу відстеження оптимуму складнішою [16]. Розглянемо застосування цих понять до задачі векторної квантизації та алгоритму K-середніх, описаних у попередньому розділі. У стандартній постановці, набір

даних $\Xi = \{\xi_1, \xi_2, \cdots, \xi_N\}$ вважається фіксованим, і метою є знаходження множини центроїдів $\{y_1, y_2, \cdots, y_K\}$, що мінімізують цільову функцію (1.7). Однак у динамічному середовищі, де дані надходять послідовно з часозалежних розподілів P_t, ця постановка потребує модифікації. Нехай у момент часу t спостережено вибірку $\Xi_t = \{\xi_{t,1}, \xi_{t,2}, \cdots, \xi_{t,m}\}$, де кожне $\xi_{t,i} \sim P_t$ є незалежною реалізацією з розподілу P_t. Тоді очікувана цільова функція векторної квантизації у момент t визначається як:

$$F_t(\{y_j\}_{j=1}^K) = \mathbb{E}_{\xi \sim P_t}\left[\min_{j=1,\ldots,K} \|\xi - y_j\|_2^2\right] \tag{1.33}$$

Відповідно, оптимальне розміщення центроїдів у момент t задається як $\{y_{t,j}^*\}_{j=1}^K = \arg\min_{\{y_j\}_{j=1}^K} F_t(\{y_j\}_{j=1}^K$. Це розміщення визначає розбиття простору ознак $\mathbb{R}^d$ методом діаграми Вороного, адаптоване до поточного розподілу P_t [7]. Проблема виникає, коли алгоритм кластеризації навчається на даних з розподілу P_s у момент часу s, а згодом застосовується до даних з іншого розподілу P_t у момент часу $t \neq s$. Розглянемо конкретний сценарій: припустимо, що центроїди $\{y_{s,j}^*\}_{j=1}^K$ були оптимально обчислені на основі даних з розподілу P_s, але розподіл з часом змістився до P_t. Якщо ці центроїди використовуються для квантизації нових даних $\xi \sim P_t$, очікувана помилка квантизації становить $F_t(\{y_{s,j}^*\}_{j=1}^K)$, тоді як оптимальна помилка для розподілу P_t дорівнює $F_t(\{y_{t,j}^*\}_{j=1}^K)$. Деградація ефективності кількісно виражається як:

$$\epsilon_{\text{drift}}(s,t) = F_t(\{y_{s,j}^*\}_{j=1}^K) - F_t(\{y_{t,j}^*\}_{j=1}^K) \geq 0 \tag{1.34}$$

де нерівність випливає з визначення оптимальності $\{y_{t,j}^*\}_{j=1}^K$ для розподілу P_t. Величина $\epsilon_{\text{drift}}(s,t)$ монотонно зростає зі збільшенням розбіжності між розподілами P_s та P_t. Емпіричні дослідження великомасштабних баз даних демонструють, що розподільний дрейф є поширеним явищем у реальних прикладних задачах.

В якості конкретного прикладу проблеми розподільного дрейфу уявімо пошуковий рушій, що індексує наукові тексти за допомогою мовної моделі DNN, що перетворює дискретну множину текстів у простір неперервних ознак [13,123]. Нехай проіндексований простір, що зображений на рисунку 1.2, розбиває множини за трьома загальними категоріями дослідження, кожна з який позначена відмінним кольором на графіку: «машинне навчання», «теорія оптимізації» та «комп'ютерна

інженерія». При додаванні статей на нові теми дослідження (позначені сірим кольором на графіку) за існуючими категоріями, наприклад, про індексацію IVF, буде зміщення розподілу тексту певної категорії змодельованого методом кластеризації, як зображено на рисунку 1.3.

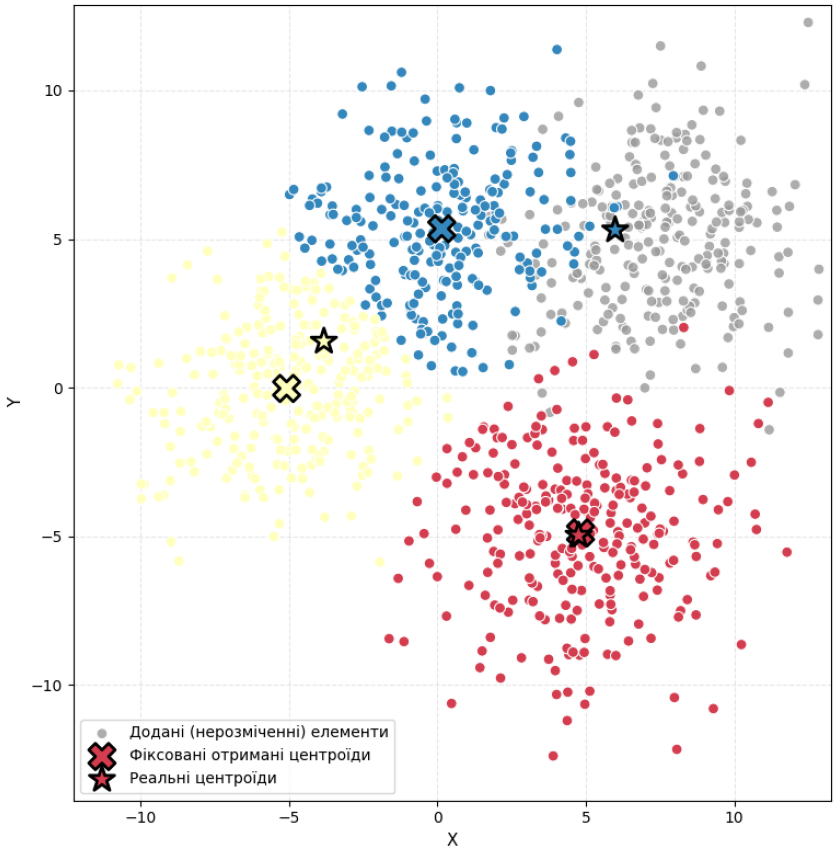

Рис 1.3: Розбиття простору ознак для трьох центроїдів методом векторної квантизації.

Це призводить до того, що реальний центроїд певного розбиття категорії буде відрізнятись від попередньо отриманого центроїду, що призводить до падіння точності моделі кластеризації. Для кількісної характеризації дрейфу у часі, вводиться поняття попарної подібності між місяцями. Нехай Φ_i та Φ_j позначають вибірки даних за місяці i та j відповідно. Асиметрична міра подібності між цими вибірками визначається як [7]:

$$S(\Phi_i, \Phi_j) = -\frac{1}{|\Phi_i|} \sum_{\phi \in \Phi_i} \mathrm{dist}(\phi, \Phi_j) \tag{1.35}$$

де $\mathrm{dist}(\phi, \Phi_j) = \frac{1}{L} \sum_{\ell=1}^{L} \|\phi - \mathrm{NN}_\ell(\phi, \Phi_j)\|_2$ є середньою відстанню від елемента ϕ до його L найближчих сусідів у множині Φ_j, а $\mathrm{NN}_\ell(\phi, \Phi_j)$ позначає ℓ-го найближчого сусіда.

Для індексних структур, що базуються на методі IVF, дрейф призводить до специфічних проблем ефективності. В IVF-індексах, простір ознак розбивається методом Вороного за допомогою методу квантизації з K центроїдами [7,49,110], а вектори бази даних зберігаються у відповідних кластерах. Під час пошуку, алгоритм відвідує лише підмножину кластерів, найближчих до запиту, що забезпечує ефективність. Проте, якщо центроїди були навчені на розподілі P_s, а поточні дані та запити походять з розподілу P_t, виникає проблема незбалансованості кластерів: деякі кластери стають переповненими векторами, які насправді ближчі до інших, змінених розподілом, оптимальних позицій центроїдів. Це призводить до збільшення кількості обчислень відстаней, необхідних для досягнення заданої точності пошуку.

Формально, нехай $\mathrm{DCS}(q,t)$ позначає кількість обчислень відстаней, необхідних для відповіді на запит $q \sim P_t$ з використанням індексу, побудованого на основі центроїдів $\{y^*_{s,j}\}_{j=1}^K$. Деградація ефективності через дрейф проявляється у тому, що $\mathbb{E}_{q\sim P_t}[\mathrm{DCS}(q,t)] > \mathbb{E}_{q\sim P_t}[\mathrm{DCS}^*(q,t)]$, де $\mathrm{DCS}^*(q,t)$ відповідає кількості обчислень відстаней для оптимально перебудованого індексу з центроїдами $\{y^*_{t,j}\}_{j=1}^K$ [7]. Більше того, деградація точності пошуку виражається у зниженні коефіцієнта повноти: частка істинних k найближчих сусідів, знайдених алгоритмом, зменшується при збільшенні розбіжності між розподілами навчання та запитів [49].

Таким чином, розподільний дрейф створює компроміс у дизайні систем індексації: з одного боку, повне перенавчання індексу на кожному часовому кроці забезпечує оптимальну продуктивність $F_t(\{y^*_{t,j}\}_{j=1}^K)$, але вимагає $O(N)$ обчислювальних операцій, де N – розмір бази даних. З іншого боку, використання статичного індексу без оновлень накопичує деградацію $\epsilon_{\mathrm{drift}}(s,t)$ з часом.

1.7 Висновки до розділу

Отже, у першому розділі було виконано постановку задачі машинного навчання «без учителя» та проведено ґрунтовний аналіз наявних методів векторної квантизації, включаючи як і класичний метод K-середніх, так і поширені похідні

методи. Детально розглянуті алгоритмічні властивості методів, а саме швидкість збіжності, стійкість до статистичних шумів тощо.

Незважаючи на такі переваги як простота програмної реалізації та гарантована монотонна збіжність, усі вищеперелічені методи спроектовані для роботи лише в умовах статичних даних, та не адаптовані до динамічних середовищ. Через феномен розподільного дрейфу, точність даних методів падає з часом у середовищах де дані мають властивість змінюватись через сезонні та соціологічні тренди, що охоплює широку кількість сучасних задач машинного навчання.

На підставі виявлених викликів було представлено запропонований альтернативний метод векторної квантизації з адаптивним оновленням центроїдів, що використовує стохастичну апроксимацію для оцінки оптимального значення центроїдів при додаванні нових елементів у набір даних $\{\xi_i\}_{i=1}^{N}$, вирішуючи проблему розподільного дрейфу в динамічних середовищах кластерного аналізу.

РОЗДІЛ 2. МЕТОДИ СТОХАСТИЧНОЇ ОПТИМІЗАЦІЇ ДЛЯ ОЦІНКИ ОПТИМАЛЬНИХ ЗНАЧЕНЬ ЦЕНТРОЇДІВ

У цьому розділі розглядаються основні відомості теорії стохастичної оптимізації, на яких базується запропонований метод SQC, а саме метод градієнтного спуску та елементи стохастичної апроксимації. Також увагу приділено модифікованим методам градієнтного спуску з адаптивним кроком, які активно використовуються для оптимізації на розріджених даних, наприклад, у задачах NLP [27,96].

Запропонований метод SQC для оцінювання оптимальних значень центроїдів спирається на оптимізацію цільової функції відстані за допомогою методів градієнтного спуску. Ключові властивості методу SQC, такі як збіжність, залежать від властивостей розглянутих у даному розділі градієнтних методів оптимізації, тому їх детальний огляд є необхідним.

Розділ починається з визначення ітеративної рекурентної послідовності, виведення її оцінки збіжності та узагальнення для задачі стохастичної оптимізації. Показано застосування методів стохастичної апроксимації Кіфера-Вольфовіца та Роббінса-Монро для оцінки оптимальних значень цільової функції на підмножині вхідних даних, що зменшує вимоги до пам'яті обчислювальних пристроїв у задачах машинного навчання. Далі розглядається міні-пакетний варіант методу для контролю дисперсії значення градієнта $\mathbb{V}\{u^k|w^k\}$. Також визначено адаптивні методи для знаходження розв'язку на розріджених даних за допомогою нормалізації значення множника кроку ρ відносно акумульованих значень градієнтів.

Зрештою, розглянуто важливі елементи адаптивних стохастичних градієнтних методів для розкриття принципу роботи запропонованого алгоритму та аналітичного доведення теоретичних гарантій збіжності. Пряма оптимізація цільової функції надає значні переваги у контролі збіжності у порівнянні з непрямою двоетапною оптимізацією сумарної функції квадратичної помилки (1.7), які будуть розглянуті в наступних розділах. Дослідження, представлені в даному розділі, були проілюстровані в публікації [92] та апробовані на конференції [89].

2.1 Огляд методів градієнтного спуску

Градієнтні методи застосовуються для ітеративного знаходження оптимального значення цільової функції зазвичай при відсутності аналітичного розв'язку або якщо функція є занадто складною для обчислення, що зробило метод популярним для оптимізації в задачах машинного навчання [96,131]. Простота реалізації, інформація про ландшафт цільової функції та масштабованість на моделях з мільярдами параметрів – ці особливості методу зробили можливими дослідження DL, такі як відомий алгоритм зворотного поширення помилки [107] та перша велика модель класифікації AlexNet [63]. Варто відмітити також їхній значний внесок у розвиток генеративного ШІ, оскільки вони дають змогу масштабування оптимізації для великих мовних моделей, що спираються на архітектуру Transformer [123].

Визначення рекурентної формули класичного методу градієнтного спуску можна отримати з аналітичним шляхом з визначення самого градієнта. Припустимо, що у кожній точці $x \in \mathbb{R}^n$ n-вимірного евклідового простору $\mathbb{R}^n$ функція $f : \mathbb{R}^n \to \mathbb{R}$ має похідну $\partial F(x,s)/\partial x$ за будь-яким напрямом $s \in \mathbb{R}^n$ [32]:

$$\frac{\partial f}{\partial x} \equiv \lim_{\rho \to +0} \frac{f(x+\rho s) - f(x)}{\rho}, \quad \nabla f = (\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n}) \tag{2.1}$$

За теоремою Тейлора [133], рухаючись з точки x в напрямку s з достатньо малим кроком $\rho > 0$ отримаємо:

$$f(x+\rho s) \approx f(x) + \rho \nabla f(x;s) + o(\rho) \tag{2.2}$$

де величина $o(\rho)$ задовольняє границі $\lim_{\rho \to 0} o(\rho)/\rho = 0$. Відповідно, для монотонного зменшення значення цільової функції напрям s має задовольняти умові:

$$\sum_{j=1}^{n} \frac{\partial f(x_j; s_j)}{\partial x_j} < 0 \tag{2.3}$$

Для диференційованої у точці x функції f відомий зв'язок між обчисленим у цій точці її градієнтом $\nabla f(x)$ і похідною $\partial f(x;s)/\partial x$ за будь-яким напрямом s:

$$\frac{\partial f(x;s)}{\partial x} = < s, \nabla f(x) > \tag{2.4}$$

Звідси, якщо напрям s утворює тупий кут з градієнтом $\nabla f(x)$, тобто якщо $< s, \nabla f(x) >< 0$ з визначення (2.4), то при будь-якому достатньо малому додатному ρ виконується нерівність $f(x + \rho s) < f(x)$. Більше того, якщо при цьому напрям s протилежний до градієнта $\nabla f(x)$, наприклад якщо $s = -\nabla f(x)$, то і швидкість зменшення значення $f(x)$ буде найбільшою, оскільки серед усіх можливих напрямів S одиничної довжини саме для вектора $-\nabla f(x)|\nabla f(x)|^{-1}$ похідна за напрямом $f'(x;\cdot)$ набуває найменшого значення:

$$\min_{s:|s|=1} f'(x;s) = \min_{s:|s|=1} < s, \nabla f(x) >= -|\nabla f(x)| \tag{2.5}$$

Очевидно, що для стаціонарних точок функції f, зокрема, для можливих розв'язків задачі $f(x) \to \inf$, нерівність $< s, \nabla f(x) >< 0$ не виконується для жодного напряму $s \in \mathbb{R}^n$. Отже, комбінуючи співвідношення (2.2) та (2.5) ми отримаємо рекурентну формулу класичного методу градієнтного спуску для ітеративного знаходження оптимального значення цільової функції $f(x)$:

$$x^{k+1} = x^k - \rho \nabla f(x^k), \quad \lim_{k\to\infty} x^k = x^* \tag{2.6}$$

У класичному методі градієнтного спуску параметр кроку залишається або фіксованим ρ або його значення визначається через числову спадаючу послідовність ρ_k. На практиці ефективність методу залежить саме від вибору «правильного» значення. Обравши занадто мале значення, ми отримаємо повільну збіжність, а обравши занадто велике значення – велику амплітуду коливань значення цільової функції або навіть розбіжність ітеративної послідовності. Метод найшвидшого спуску на кожній ітерації визначає оптимальний розмір кроку ρ_k шляхом розв'язання допоміжної задачі одновимірної оптимізації. Зокрема, на кожному кроці k необхідно знайти точку мінімуму ρ_k функції однієї змінної [32]:

$$\theta(\rho_k) \equiv f(x^k - \rho_k \nabla f(x^k)) \to \min_{\rho_k > 0} \tag{2.7}$$

Після знаходження оптимального значення ρ_k наступне наближення оптимального значення x^{k+1} знаходиться за співвідношенням (2.6).

Для аналізу збіжності методу градієнтного спуску необхідно встановити умови на цільову функцію f та правило вибору кроку ρ. Розглянемо випадок, коли функція f є опуклою та двічі неперервно диференційованою, а її матриця Гессе

$\nabla^2 f(x)$ задовольняє умову обмеженості при певних додатних константах m і M, $m \le M$ [133]:

$$m\|y\|^2 \le \langle y, \nabla^2 f(x) y\rangle \le M\|y\|^2, \quad \forall x, y \in \mathbb{R}^n \tag{2.8}$$

Ця умова означає, що функція f є m-сильно опуклою та має M-ліпшицевий градієнт. За таких умов, якщо множина рівня $\{x \in \mathbb{R}^n : f(x) \le f(x^0)\}$ є обмеженою, задача оптимізації $f(x) \to \inf\limits_x$ має єдиний розв'язок x^* [124].

Критерій збіжності методу суттєво залежить від вибору кроку ρ у рекурентному співвідношенні (2.6). Практично ефективним є правило Арміхо, яке полягає у виборі найбільшого значення ρ з множини $\{\bar{\rho}, 2^{-1}\bar{\rho}, 2^{-2}\bar{\rho}, ...\}$, для якого виконується умова достатнього спуску:

$$f(x^{k+1}) \le f(x^k) - \frac{\varepsilon\rho}{2}\|\nabla f(x^k)\|^2 \tag{2.9}$$

де $\varepsilon \in (0,1)$ – фіксований параметр. При застосуванні правила Арміхо гарантується нижня межа кроку $\rho \ge 1/(2M)$.

Теорема 1 (Збіжність градієнтного спуску) [124] Нехай X – гільбертів простір, $f : X \to \mathbb{R}$ – опукла функція з M-ліпшицевим градієнтом та m-сильно опукла з $m \ge 0$ [82]. Нехай $x^0 \in X$ – початкове наближення, а крок $\rho_k > 0$ визначається правилом Арміхо (2.9). Тоді для методу мінімізації $x^* = \arg\min_{x \in X} f(x)$ виконуються оцінки:

$$f(x^n) - f(x^*) \le \frac{M\|x^0 - x^*\|^2}{n}, \quad \|x^n - x^*\|^2 \le \left(1 - \frac{m}{2M}\right)^n \|x^0 - x^*\|^2 \tag{2.10}$$

З теореми випливає, що при $m > 0$ (сильно опукла функція) метод збігається з лінійною швидкістю, тобто $\|x^{k+1} - x^*\| \le \gamma\|x^k - x^*\|$ для $\gamma = \sqrt{1 - m/(2M)} < 1$. У загальному випадку опуклої функції ($m = 0$) швидкість збіжності є сублінійною: $f(x^n) - f(x^*) = O(1/n)$.

Класичний градієнтний спуск широко застосовується у задачах машинного навчання завдяки простоті реалізації та низьким вимогам до пам'яті обчислювального пристрою. Проте метод має суттєві обмеження: повільна збіжність на погано обумовлених задачах (коли відношення M/m є великим), чутливість до вибору початкового наближення та труднощі з проходженням сідлових точок і плато цільової функції [100].

Метод «важкої кулі», запропонований Б.Т.Поляком [98,99], використовує фізичну аналогію руху тіла в потенціальному полі під дією сили тертя для подолання обмежень класичного градієнтного спуску. Ідея полягає у введенні множника прискорення γ, що дозволяє накопичувати інформацію про попередні напрямки руху. Поведінка градієнтного спуску моделюється ньютонівським диференціальним рівнянням другого порядку:

$$a\frac{\partial^2 x}{\partial t^2} + v\frac{\partial x}{\partial t} = -\nabla_x f(x) \tag{2.11}$$

де $a > 0$ – множник маси, $v > 0$ – коефіцієнт тертя, а градієнт $\nabla_x f(x)$ представляє консервативне силове поле з потенціальною енергією $f(x)$ [98]. Застосовуючи скінченно-різницеву апроксимацію похідних з кроком Δt, отримаємо:

$$a\frac{x^{k+1} - 2x^k + x^{k-1}}{\Delta t^2} + v\frac{x^{k+1} - x^k}{\Delta t} = -\nabla_x f(x^k) \tag{2.12}$$

Перегрупувавши члени рівняння та ввівши позначення $\rho = \Delta t^2/(a + v\Delta t)$ для кроку градієнтного спуску і $\gamma = a/(a + v\Delta t)$ для множника імпульсу, отримаємо рекурентну формулу методу імпульсу:

$$x^{k+1} = x^k + \gamma_k(x^k - x^{k-1}) - \rho_k \nabla f(x^k), \quad x^0 \in \mathbb{R}^n, \quad k \in \mathbb{N} \tag{2.13}$$

Множники $\gamma_k > 0$ та $\rho_k > 0$ можуть залежати від поточної точки x^k або всієї історії $\{x^0, \dots, x^k\}$. Замість детермінованих градієнтів $\nabla f(x^k)$ можна використовувати стохастичні градієнти, що робить метод придатним для задач стохастичної оптимізації. Покажемо, що побудована рекурентна послідовність є збіжною та виведемо верхню оцінку швидкості збіжності.

Теорема 2 (Збіжність методу «важкої кулі») [99] Нехай $f : H \to \mathbb{R}$ – двічі диференційована функція в гільбертовому просторі H у деякому околі точки мінімуму $x^* = \arg\min_{x \in H} f(x)$, причому гессіан $\nabla^2 f(x^*)$ є самоспряженим оператором, що задовольняє для всіх $y \in H$ умові (2.9). Нехай $x^0,\ x^1 \in H$ – початкові наближення, достатньо близькі до x^*, і послідовність $\{x^n\}$ визначається методом імпульсу (2.13) з оптимальними параметрами:

$$\rho = \frac{4}{(\sqrt{M} + \sqrt{m})^2}, \quad \gamma = \left(\frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}}\right)^2 \tag{2.14}$$

Тоді послідовність збігається до x^* зі швидкістю геометричної прогресії: для довільного $\varepsilon > 0$ існує стала $C(\varepsilon) > 0$ така, що

$$f(x^n) - f(x^*) \le \frac{M\,C(\varepsilon)}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}+\varepsilon\right)^{2n}(\|x^0-x^*\|^2+\|x^1-x^*\|^2) \tag{2.15}$$

де $\kappa = M/m$ – число обумовленості гессіана. Оцінка (2.15) випливає з $f(x)-f(x^*) \le (M/2)\cdot\|x-x^*\|^2$ та основної оцінки на відстань

$$\|x^n - x^*\| \le C(\varepsilon)\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}+\varepsilon\right)^{n}(\|x^0-x^*\|^2+\|x^1-x^*\|^2)^{1/2} \tag{2.16}$$

Для квадратичного функціоналу збіжність (2.15) має місце глобально для довільних $x^0, x^1 \in H$. Для порівняння, метод градієнтного спуску (2.13) при $\gamma = 0$ з оптимальним кроком $\rho = 2/(M+m)$ дає суттєво повільнішу оцінку:

$$f(x^n) - f(x^*) \le \frac{M}{2}\left(\frac{\kappa-1}{\kappa+1}\right)^{2n}\|x^0-x^*\|^2 \tag{2.17}$$

оскільки для великих κ виконується

$$\frac{\kappa-1}{\kappa+1} \approx 1-\frac{2}{\kappa} \gg \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \approx 1-\frac{2}{\sqrt{\kappa}} \tag{2.18}$$

Імпульсний множник прискорює збіжність алгоритму в областях плато та сідлових точках цільової функції, де градієнт є малим за величиною, як продемонстровано на рисунку 2.1.

Фізична інтерпретація полягає в тому, що «важка куля» накопичує швидкість при русі вниз по схилу потенціальної поверхні і здатна долати невеликі підйоми за рахунок інерції [99]. Проте накопичення імпульсу ускладнює контроль збіжності в околі мінімуму: алгоритм може «перестрибувати» оптимальну точку і вимагати додаткових ітерацій для зменшення осциляцій. Ця проблема особливо виражена при великих значеннях γ або на функціях з вузькими ярами.

Прискорений градієнтний метод Нестерова (або метод яружного кроку) [81], вирішує проблему осциляцій методу імпульсу за рахунок екстраполяції при обчисленні градієнта. Ключова ідея полягає в тому, що градієнт обчислюється не в поточній точці , а в екстрапольованій точці , яка враховує напрямок попереднього руху.

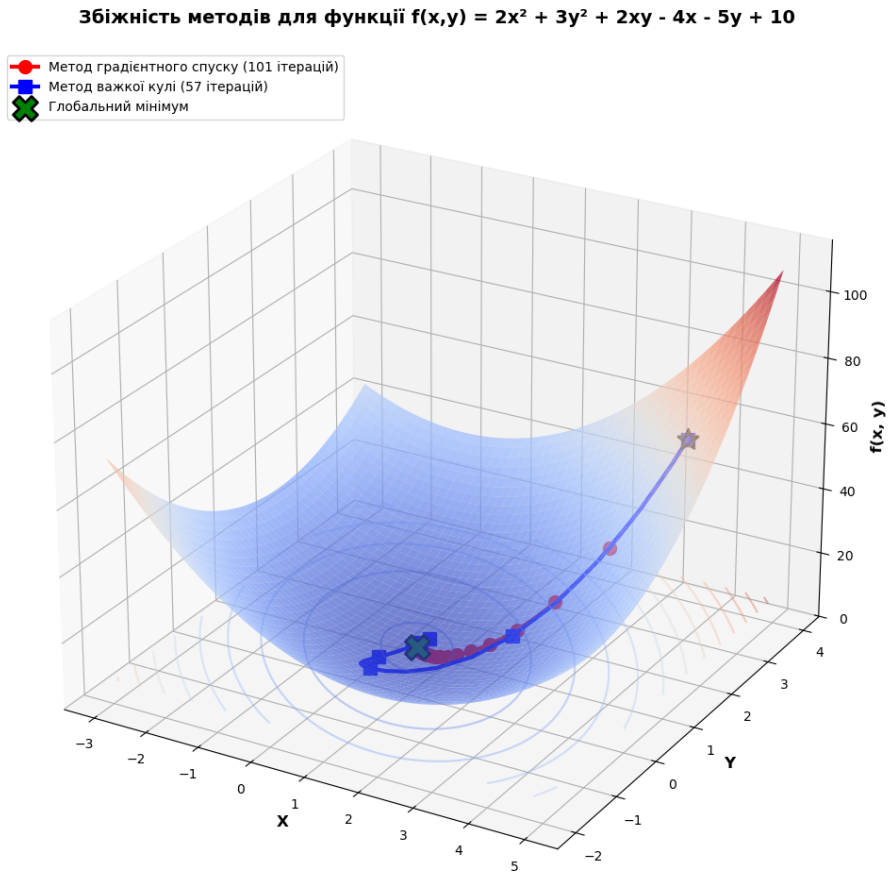

Рис. 2.1: Порівняння швидкості збіжності методу градієнтного спуску та методу «важкої кулі» при знаходженні мінімуму функції $f(x, y) = 2x^2 + 3y^2 + 2xy - 4x - 5y + 10$ при заданих параметрах $\rho = 0.01$, $\gamma = 0.8$ та початковій точці $(4.0, 3.0)$. Незважаючи на необхідність в додаткових ітераціях для погашення імпульсу, метод «важкої кулі» досяг мінімуму швидше ніж класичний метод.

Метод визначається наступною рекурентною послідовністю [81,124]:

$$\begin{aligned} x^{k+1} &= y^k - \rho_k \nabla f(y^k) \\ y^{k+1} &= x^{k+1} + \tfrac{\lambda_k - 1}{\lambda_{k+1}}(x^{k+1} - x^k), \\ x^0 &= y^0 \in \mathbb{R}^n, \quad k = 0, 1, 2, \ldots \end{aligned} \tag{2.19}$$

де $\rho_k > 0$ – розмір кроку, а послідовність $\{\lambda_k\}_{k=0}^{\infty}$ визначається рекурентно:

$$\lambda_0 = 0, \quad \lambda_{k+1} = \frac{1 + \sqrt{1 + 4\lambda_k^2}}{2}, \quad k = 0, 1, 2, \ldots \tag{2.20}$$

Перше рівняння методу (2.19) означає градієнтний крок з екстрапольованої точки y^k до нової точки x^{k+1}, а друге рівняння – крок вздовж напрямку $x^{k+1} - x^k$ з

коефіцієнтом $(\lambda_k - 1)/\lambda_{k+1}$. Послідовність $\{\lambda_k\}$ задовольняє умову $\lambda_k(\lambda_k - 1) \leq \lambda_{k-1}^2$ з рівністю, що забезпечує оптимальність вибору параметрів [124].

Теорема 3 (Збіжність методу Нестерова) [124] Нехай X – гільбертів простір, $f : X \to \mathbb{R}$ – опукла функція з M-ліпшицевим градієнтом. Припустимо, що $x^0 = y^0 \in X$, а крок $\rho_k > 0$ визначається правилом Арміхо з додатковою умовою $\rho_k \leq \rho_{k-1}$. Тоді для методу Нестерова $x^* = \arg\min_{x \in X} f(x)$ виконуються оцінки:

$$f(x^n) - f(x^*) \leq \frac{M\|x^0 - x^*\|^2}{\lambda_{n-1}^2} \leq \frac{4M\|x^0 - x^*\|^2}{(n+1)^2} \tag{2.21}$$

оскільки для послідовності (2.21) виконується $\lambda_{n-1} \geq (n-1)/2$.

Ця швидкість збіжності $O(1/n^2)$ є суттєвим покращенням порівняно зі швидкістю $O(1/n)$ класичного градієнтного спуску. Більше того, доведено, що ця швидкість є оптимальною для класу методів першого порядку на опуклих функціях з ліпшицевим градієнтом. Для визначення швидкості збіжності отриманого методу (2.19) на опуклих функціях спочатку дамо визначення даному класу функцій.

Опукла функція $f : X \to \mathbb{R}$, визначена на опуклій множині $X \subset \mathbb{R}^n$, називається m-сильно опуклою з параметром $m > 0$, якщо для всіх $x, y \in X$ та $t \in [0, 1]$ виконується нерівність [82]:

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \frac{m}{2}t(1-t)\|x - y\|^2 \tag{2.22}$$

Інтуїтивно, сильно опукла функція зростає не повільніше за квадратичну функцію. Зазначимо, що при $m = 0$ визначення (2.22) збігається з класичним визначенням опуклої функції, а при $m \to 0^+$ воно наближається до визначення строго опуклої функції. Проте існують функції, що є строго опуклими, але не є сильно опуклими для жодного $m > 0$: наприклад, якщо для строго опуклої функції f існує послідовність точок (x_n) така, що $f''(x_n) = 1/n$, то незважаючи на виконання умови $f''(x_n) > 0$, функція не є сильно опуклою, оскільки $f''(x)$ стає довільно малим [11]. Для диференційованої функції f еквівалентним є наступне визначення сильної опуклості з параметром $m > 0$: для всіх $x, y \in X$ має виконуватись нерівність [82]:

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq m\|x - y\|^2 \tag{2.23}$$

де $\langle \cdot, \cdot \rangle$ позначає внутрішній добуток, а $\| \cdot \|$ – відповідну норму. Ця умова означає, що градієнт ∇f є m-сильно монотонним оператором. Еквівалентною є також умова на нижню оцінку функції через її лінійну апроксимацію [11]:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2}\|y - x\|^2, \quad \forall x, y \in X \tag{2.24}$$

Геометрично це означає, що графік функції f лежить не нижче параболоїда з кривизною m, дотичного до графіка в точці x. Для двічі неперервно диференційовних функцій існує зручний критерій сильної опуклості через матрицю Гессе. Функція f є m-сильно опуклою тоді і тільки тоді, коли для всіх $x \in X$ виконується умова [82]:

$$\nabla^2 f(x) \succeq mI \tag{2.25}$$

де I – одинична матриця, а нерівність $\succeq$ означає, що матриця $\nabla^2 f(x) - mI$ є додатно напіввизначеною. Аналогічно, мінімальне власне значення матриці Гессе $\lambda_{\min}(\nabla^2 f(x))$ має бути не меншим за m для всіх $x \in X$. У одновимірному випадку ця умова набуває вигляду $f''(x) \geq m$ для всіх x в області визначення. Доведення імплікації від умови (2.25) до (2.24) базується на теоремі Тейлора [133], яка стверджує, що існує точка $z \in \{tx + (1-t)y : t \in [0,1]\}$ така, що:

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x) \tag{2.26}$$

З припущення про власні значення матриці Гессе випливає $(y-x)^T \nabla^2 f(z)(y-x) \geq m\|y-x\|^2$, що дає нерівність (2.24) [11]. Важливою характеризацією сильної опуклості є наступне твердження: функція f є m-сильно опуклою тоді і тільки тоді, коли функція $g(x) = f(x) - \frac{m}{2}\|x\|^2$ є опуклою [82]. Ця властивість дозволяє зводити аналіз сильно опуклих функцій до аналізу опуклих функцій.

Сильно опуклі функції мають ряд важливих властивостей, що спрощують аналіз оптимізаційних задач [11,82]: для кожного дійсного числа r множина рівня $\{x \in X : f(x) \leq r\}$ є компактною та функція f має єдиний глобальний мінімум на $\mathbb{R}^n$. Єдиність мінімуму випливає безпосередньо з нерівності (2.22): якби існували дві різні точки мінімуму x^* та y^*, то для $t = 1/2$ отримали б суперечність $f((x^* + y^*)/2) < (1/2)f(x^*) + (1/2)f(y^*) = f(x^*)$. Використовуючи отримані нерівності

(2.24)-(2.25) та властивості сильно опуклих функцій вище, ми можемо показати, що для сильно опуклих функцій з параметром $m > 0$ можливе досягнення лінійної швидкості збіжності при фіксованому виборі параметра λ.

Теорема 4 (Збіжність методу Нестерова для сильно опуклої функції) [124] Нехай $f : X \to \mathbb{R}$ є m-сильно опуклою з M-ліпшицевим градієнтом, і нехай $\lambda \geq \sqrt{2M/m} > 1$ – фіксований параметр. Тоді для методу Нестерова з фіксованим параметром виконується:

$$f(x^n) - f(x^*) \leq \left(1 - \frac{1}{\lambda}\right)^n (2\rho_0 \lambda^2 (f(x^0) - f(x^*)) + \|x^0 - x^*\|^2) \cdot \frac{M}{\lambda^2} \tag{2.27}$$

При оптимальному виборі $\lambda = \sqrt{M/m}$ константа збіжності становить $1 - 1/\sqrt{\kappa}$, де $\kappa = M/m$ – число обумовленості задачі. Це суттєво краще за константу $1 - 1/\kappa$ класичного градієнтного спуску для погано обумовлених задач.

2.2 Стохастична апроксимація

Розглянуті методи оптимізації застосовуються переважно для задач з детермінованою цільовою функцією $f(x)$, де власне функція та її градієнт $\nabla f(x)$ можуть бути обчислені точно для будь-якого значення аргументу x. Однак для задач стохастичної оптимізації, що задаються рівнянням (1.30), дані методи мають ряд обмежень оскільки безпосереднє обчислення градієнта є неможливим або трудомістким, оскільки розподіл даних може змінюватись з часом і цільова функція $f(x) = \mathbb{E}\, F(x, \xi)$ визначається неявно. Для задачі (1.30) якщо функція є диференційованою, її градієнт також визначається як математичне сподівання:

$$\nabla f(x) = \mathbb{E}_\xi [\nabla_x F(x, \xi)] = \int_{\xi \in \Xi} \nabla_x F(x, \xi)\, P(d\xi) \tag{2.28}$$

Оскільки значення градієнта є інтегралом за розподілом даних, то замість точного обчислення градієнта використовувати його статистичну оцінку за скінченною випадковою вибіркою, тобто замінити математичне сподівання його аналогом за методом Монте-Карло [14]. Коли розмір вибірки $N_k = N$ і використовується весь набір даних, ми отримуємо класичний градієнтний спуск на емпіричному ризику. Коли $N_k = 1$, маємо граничний випадок – стохастичний градієнтний спуск з

максимальним рівнем шуму, але мінімальною обчислювальною вартістю на ітерацію.

Теоретичні основи методів стохастичної апроксимації були закладені у початкових роботах Роббінса та Монро [104] та Кіфера та Вольфовіца [55]. Ці роботи запропонували рекурентні процедури для знаходження коренів та екстремумів регресійних функцій за наявності випадкового шуму і заклали основу для сучасного методу стохастичного градієнтного спуску. Розглянемо спочатку задачу знаходження кореня θ рівняння $M(x) = \alpha$, де α – задана стала, а $M(x)$ – невідома регресійна функція, визначена як [104]:

$$M(x) = \int_{-\infty}^{\infty} y \, dP(y \mid x) \tag{2.29}$$

Тут $P(y \mid x)$ – умовна функція розподілу спостережень при заданому рівні x. Безпосереднє обчислення $M(x)$ є неможливим, проте для будь-якого значення x дослідник може отримати незалежне випадкове спостереження y з розподілом $P(y \mid x)$, для якого $\mathbb{E}[y \mid x] = M(x)$. Метод стохастичної апроксимації Роббінса-Монро визначається рекурентним співвідношенням [104]:

$$x_{n+1} - x_n = a_n(\alpha - y_n) \tag{2.30}$$

де y_n – випадкове спостереження з розподілом $P(y \mid x)$, x_1 – довільне початкове наближення, а $\{a_n\}$ – послідовність додатних множників кроку. Інтуїтивний зміст рекурсії (2.30) полягає у наступному: якщо спостережене значення $y_n > \alpha$, то $M(x_n)$, ймовірно, перевищує α, і за монотонності M наступне наближення x_{n+1} зсувається в напрямку зменшення $M(x)$; аналогічно, якщо $y_n < \alpha$, наближення рухається в бік зростання $M(x)$. Послідовність множників кроку $\{a_n\}$ називається послідовністю типу $1/n$, якщо існують додатні сталі c' та c'' такі, що $c'/n \le a_n \le c''/n$. Зокрема, для таких послідовностей виконуються умови [104]:

$$0 < \sum_{n=1}^{\infty} a_n^2 = A < \infty, \quad \sum_{n=1}^{\infty} a_n = \infty \tag{2.31}$$

Перша умова забезпечує згасання кумулятивного шуму, а друга – достатній сумарний прогрес у напрямку розв'язку. Наведемо основний результат про збіжність методу.

Теорема 5 (Збіжність методу Роббінса-Монро) [104] Нехай $M(x)$ – неспадна функція, $M(\theta)=\alpha$ та $M'(\theta)>0$. Нехай існує стала $C>0$ така, що для всіх x виконується:

$$\Pr[\,|Y(x)|\le C\,]=\int_{-C}^{C} dP(y\mid x)=1 \tag{2.32}$$

Нехай послідовність $\{a_n\}$ є типу $1/n$ і задовольняє умовам (2.31). Тоді для послідовності $\{x_n\}$, визначеної рекурсією (2.30), виконується:

$$\lim_{n\to\infty} b_n=\lim_{n\to\infty} E(x_n-\theta)^2=0 \tag{2.33}$$

тобто x_n збігається до θ в середньому квадратичному.

Метод Кіфера-Вольфовіця узагальнив ідею стохастичної апроксимації на задачу знаходження точки максимуму θ невідомої регресійної функції $M(x)$, яка є строго зростаючою при $x<\theta$ та строго спадною при $x>\theta$ [55]. На відміну від методу Роббінса-Монро, значення $M(\theta)$ тут невідоме, і задача полягає в оцінюванні похідної $M'(x)$ за допомогою скінченних різниць зашумлених спостережень. Метод Кіфера-Вольфовіца визначається рекурентною формулою:

$$z_{n+1}=z_n+a_n\,\frac{y_{2n}-y_{2n-1}}{c_n} \tag{2.34}$$

де y_{2n-1} та y_{2n} – незалежні спостереження з розподілами $H(y\mid z_n-c_n)$ та $H(y\mid z_n+c_n)$ відповідно, z_1 – довільне початкове наближення, а послідовності $\{a_n\}$ та $\{c_n\}$ задовольняють умовам:

$$c_n\to 0,\quad \sum a_n=\infty,\quad \sum a_n c_n<\infty,\quad \sum a_n^2 c_n^{-2}<\infty \tag{2.35}$$

Зокрема, вибір $a_n=n^{-1}$ та $c_n=n^{-1/3}$ задовольняє всім цим умовам. Умовне математичне сподівання різниці спостережень дорівнює $E[y_{2n}-y_{2n-1}\mid z_n]=M(z_n+c_n)-M(z_n-c_n)$, що наближує $2c_nM'(z_n)$ при малих c_n. Таким чином, алгоритм рухається в напрямку зростання M, використовуючи скінченно-різницеву оцінку градієнта. На регресійну функцію $M(x)$ накладаються наступні умови регулярності:

1. Існують додатні β та B такі, що при $|x'-\theta|+|x''-\theta|<\beta$ виконується $|M(x')-M(x'')|<B\,|x'-x''|$;
2. Існують додатні ρ та R такі, що при $|x'-x''|<\rho$ виконується $|M(x')-M(x'')|<R$;

3. Для кожного $\delta > 0$ існує $\pi(\delta) > 0$ таке, що при $|z - \theta| > \delta$ виконується:

$$\inf_{\frac{1}{2}\delta > \epsilon > 0} \frac{|M(z+\epsilon) - M(z-\epsilon)|}{\epsilon} > \pi(\delta) \tag{2.36}$$

Крім того, передбачається обмеженість другого моменту спостережень для всіх x є справедливим:.

$$\int_{-\infty}^{\infty} (y - M(x))^2 \, dP(y \mid x) \leq S < \infty \tag{2.37}$$

Теорема 6 (Збіжність методу Кіфера-Вольфовіца) [55] Нехай $M(x)$ – регресійна функція, строго зростаюча при $x < \theta$ та строго спадна при $x > \theta$, що задовольняє умовам регулярності (2.36). Нехай послідовності $\{a_n\}$ та $\{c_n\}$ задовольняють умовам (2.35). Тоді послідовність $\{z_n\}$, визначена рекурсією (2.34), збігається до θ стохастично:

$$\forall\, \eta > 0,\ \varepsilon > 0\ \ \exists N(\eta, \varepsilon)\ \ P\{|z_n - \theta| > \eta\} \leq \varepsilon\ \ n > N(\eta, \varepsilon) \tag{2.38}$$

Методи Роббінса-Монро та Кіфера-Вольфовіца становлять теоретичну основу для стохастичного градієнтного спуску. Зв'язок між ними є безпосереднім: задача мінімізації $f(x) = \mathbb{E}\, F(x, \xi)$ еквівалентна задачі знаходження кореня рівняння $\nabla f(x) = 0$, яке має вигляд $\mathbb{E}[\nabla_x F(x, \xi)] = 0$. Застосовуючи схему Роббінса-Монро до цього рівняння, ми отримуємо стохастичний градієнтний спуск.

Визначення 1 Випадковий вектор u^k називається стохастичним градієнтом функції $f(x)$ у точці $x = x^k$, якщо виконується умова [32]:

$$\mathbb{E}\{u^k \mid x^k\} = \nabla f(x^k) \tag{2.39}$$

де $\mathbb{E}\{u^k \mid x^k\}$ позначає умовне математичне сподівання.

Отже, якщо $\nabla f(x) = \mathbb{E}[\nabla_x F(x, \xi)]$, то вектор $u^k = \nabla_x F(x^k, \xi^k)$, де ξ^k – незалежна реалізація випадкової величини ξ, дійсно є стохастичним градієнтом функції $f(x)$ у точці $x = x^k$. Відповідний метод стохастичного градієнтного спуску визначається рекурентним рівнянням:

$$x^{k+1} = \Pi_X(x^k - \rho_k\, u^k), \quad x^0 \in X, \quad k \in \mathbb{N} \tag{2.40}$$

де Π_X – оператор проекції на допустиму множину X, а $\rho_k \geq 0$ – множник кроку спуску. Для забезпечення збіжності методу (2.40) множники кроку повинні задовольняти класичним умовам Роббінса-Монро [104]:

$$\sum_{k=0}^{\infty} \rho_k = +\infty, \quad \sum_{k=0}^{\infty} \rho_k^2 < +\infty \tag{2.41}$$

Ці умови стверджують, що в той час як сума множників кроків повинна прямувати до нескінченності (забезпечуючи досяжність будь-якої точки простору), сума їх квадратів повинна залишатися скінченною (забезпечуючи згасання кумулятивного статистичного шуму). Зокрема, послідовність $\rho_k = c/k$ для будь-якої сталої $c > 0$ задовольняє обом умовам. В роботі [32] також було розширено метод (2.41) на задачі з обмеженнями та з недиференційовними функціями.

Суттєвою практичною проблемою стохастичного градієнтного спуску є осциляції ітеративної послідовності навколо оптимуму, спричинені ненульовою дисперсією стохастичного градієнта. Навіть при виконанні умов (2.41), що гарантують теоретичну збіжність, на практиці спадні множники кроку суттєво уповільнюють алгоритм на фінальних ітераціях. При фіксованому кроці $\rho_k = \rho$ осциляції в околі мінімуму з амплітудою є пропорційною добутку кроку та стандартного відхилення стохастичного градієнта [100]. Саме цей ефект робить добре відомим емпіричне спостереження про повільну асимптотичну швидкість збіжності методу (2.40), обумовлену його внутрішньою дисперсією.

Ефективним способом контролю дисперсії є метод міні-пакетного градієнтного спуску. Замість обчислення стохастичного градієнта за одним спостереженням, на кожній ітерації k обирається випадкова підмножина $\mathcal{B}_k$ з $b = |\mathcal{B}_k|$ елементів, і градієнт оцінюється як середнє [100]:

$$g_b^k = \frac{1}{b} \sum_{i \in \mathcal{B}_k} \nabla_x F(x^k, \xi_i^k) \tag{2.42}$$

Оцінка g_b^k залишається незміщеною: $\mathbb{E}[g_b^k \mid x^k] = \nabla f(x^k)$. Для аналізу впливу розміру міні-пакету на дисперсію стохастичного градієнта розглянемо задачу лінійної регресії з функцією втрат $L(w) = (1/n)\sum_{i=1}^{n} L_i(w)$, де $L_i(w) = (1/2)(w^T x_i - y_i)^2$, та міні-пакетний SGD з оновленням $w_{t+1}^b = w_t^b - \alpha_t g_t^b$. Позначимо $c_b = (n-b)/(b(n-1)) \geq 0$ та $\mathcal{F}_t^b$ – оператор вибірки до t-ї ітерації.

Теорема 7 (Дисперсія міні-пакетного стохастичного градієнта) [100] Для фіксованих початкових ваг w_0 та довільної квадратної матриці $B \in \mathbb{R}^{p \times p}$, обидві величини $\mathbb{V}(Bg_t^b \mid \mathcal{F}_0)$ та $\mathbb{V}(B\nabla L(w_t^b) \mid \mathcal{F}_0)$ є спадними функціями розміру міні-пакету

b для всіх $b \in [n]$, $t \in \mathbb{N}$ та всіх виборів множників кроку, незалежних від b. Зокрема, умовна дисперсія міні-пакетного оцінювача має вигляд:

$$\mathbb{V}(Bg_t^b \mid \mathcal{F}_t^b) = c_b \left(\frac{1}{n} \sum_{i=1}^{n} \|B\nabla L_i(w_t^b)\|^2 - \|B\nabla L(w_t^b)\|^2 \right) \tag{2.43}$$

де $c_b = (n-b)/(b(n-1))$ є спадною функцією b.

Формула (2.43) демонструє, що дисперсія стохастичного оцінювача градієнта пропорційна коефіцієнту c_b, який монотонно спадає при зростанні розміру міні-пакету від $b = 1$ (максимальна дисперсія, $c_1 = 1$) до $b = n$ (нульова дисперсія, $c_n = 0$). Для глибших моделей аналітичний вигляд дисперсії є більш складним. Для двошарових лінійних мереж із L_2-функцією втрат та вибірками зі стандартного нормального розподілу було показано, що дисперсія стохастичного оцінювача градієнта є поліномом від $1/b$:

$$\mathbb{V}(g_{t,i}^b \mid \mathcal{F}_0) = \beta_1 \frac{1}{b} + \beta_2 \frac{1}{b^2} + \cdots + \beta_r \frac{1}{b^r}, \qquad r \leq 2 \cdot 3^{t+1} \tag{2.44}$$

де $\beta_1, \cdots, \beta_r$ – сталі, незалежні від b, а $\beta_1 \geq 0$. Відсутність вільного члена у формулі (2.44) узгоджується з інтуїтивним очікуванням: при $b \to \infty$ дисперсія стохастичного градієнта прямує до нуля. Невід'ємність провідного коефіцієнта β_1 гарантує існування порогового значення b_0 такого, що для всіх $b \geq b_0$ дисперсія є спадною функцією b. Емпіричні дослідження підтверджують, що поріг b_0 на практиці є малим (порядку $b_0 = 4$ для типових задач), а спадна залежність дисперсії від розміру підмножини $\mathcal{B}_k$ зберігається для DL мереж з різними функціями активації та функціями втрат.

Таким чином, розмір міні-пакету b відіграє роль регулятора між обчислювальною вартістю та стабільністю алгоритму. Збільшення b зменшує дисперсію приблизно як $O(1/b)$, що дозволяє використовувати більші ефективні кроки або досягати стабільнішої збіжності при фіксованому кроці. Проте це відбувається ціною збільшення обчислювальних витрат на кожну ітерацію, що потребує ретельного вибору b залежно від конкретної задачі та доступних обчислювальних ресурсів.

2.3 Модифікації з адаптивним кроком

Отже, метод «важкої кулі» Поляка (2.13) та метод Нестерова (2.19) суттєво покращують збіжність порівняно з класичним градієнтним спуском (2.6) за рахунок введення множника прискорення, який адаптує швидкість збіжності відповідно до ландшафту цільової функції. Показовий приріст в швидкості збіжності був продемонстрований в дослідженнях в оптимізації з використанням класичних моделей машинного навчання з відносно невеликою розмірністю [119]. Проте усі вище-перелічені методи все ще зберігають спільне обмеження, а саме застосування єдиного скалярного кроку ρ_k до всіх компонент вектора при оновленні параметрів з рекурсивним рівнянням (2.6). У сучасних прикладних задачах машинного навчання вхідні дані мають дуже високу розмірність, проте в кожному векторі ознак певного елемента лише невелика частина ознак є ненульовою [27,63,96,119] і рідкісні ознаки часто несуть значну дискримінантну інформацію, що є частою характеристикою задач NLP. Застосування однакового кроку до всіх параметрів призводить до того, що рідкісні, але інформативні ознаки оновлюються надто повільно, тоді як часті ознаки отримують надлишкові оновлення [96]. Ця проблема мотивує розробку адаптивних методів, які динамічно підлаштовують крок навчання для кожного параметра окремо на основі інформації про кожну компоненту градієнта [27,56,92,121]. Ідея адаптації виходить з аналізу кумулятивної структури градієнтної інформації. Згідно з умовами збіжності рекурентної послідовності градієнтного спуску (2.40), при кількості ітерацій, що наближається до нескінченності, сума градієнтів вказує в бік мінімуму. Хоча точне значення цього «ідеального» напрямку недосяжне при побудові ітераційної послідовності, його наближення можна отримати, використовуючи накопичену евклідову норму послідовності градієнтів [27]. Саме цей принцип лежить в основі сімейства адаптивних методів, що розглядаються нижче.

Метод AdaGrad, реалізує ідею покоординатної адаптації кроку на основі історії спостережених субградієнтів. Ключове спостереження полягає в тому, що для кожної координати j параметричного простору оптимальний крок оновлення параметрів визначається як розв'язок наступної задачі мінімізації [27]:

$$\min_{s} \sum_{t=1}^{T} \sum_{j=1}^{n} \frac{(g_j^t)^2}{s_j}, \quad s \succeq 0, \quad \langle \mathbf{1}, s \rangle \leq c \tag{2.45}$$

де g_j^t позначає j-ту компоненту субградієнта на ітерації t, а $c > 0$ є обмеженням на суму компонент. Задача (2.45) розв'язується шляхом побудови лагранжіана та аналізу умов оптимальності. Вводячи множники Лагранжа $\lambda \succeq 0$ та $\theta \geq 0$, отримаємо [27]:

$$\mathcal{L}(s, \lambda, \theta) = \sum_{j=1}^{n} \frac{\|g_{1:T,j}\|_2^2}{s_j} - \langle \lambda, s \rangle + \theta(\langle \mathbf{1}, s \rangle - c) \tag{2.46}$$

де $g_{1:T,j} = [g_j^1, \dots, g_j^T]$ є j-ю лінійною комбінацією субградієнтів за всі ітерації. Обчислення часткових похідних дають оптимальний розв'язок $s_j = c \, \|g_{1:T,j}\|_2 / \sum_{i=1}^{n} \|g_{1:T,i}\|_2$. Підставляючи цей розв'язок назад у цільову функцію (2.46), отримаємо:

$$\inf_{s} \left\{ \sum_{t=1}^{T} \sum_{j=1}^{n} \frac{(g_j^t)^2}{s_j} : s \succeq 0, \, \langle \mathbf{1}, s \rangle \leq c \right\} = \frac{1}{c} \left(\sum_{j=1}^{n} \|g_{1:T,j}\|_2 \right)^2 \tag{2.47}$$

Цей результат мотивує використання функції $\psi_t(x) = \langle x, H_t x \rangle$ з діагональною матрицею $H_t = \delta I + \operatorname{diag}(s_t)$, де $s_{t,j} = \|g_{1:t,j}\|_2$ і $\delta \geq 0$ є параметром регуляризації. Позначимо $G_k = G_{k-1} + (g_j^k)^2$ накопичену суму квадратів градієнтів, тобто $G_{k,j} = \sum_{\tau=1}^{k} (g_j^\tau)^2 = \|g_{1:k,j}\|_2^2$. Визначивши адаптивний крок як $\bar{\rho}_k = \rho_k / \sqrt{G_k + \varepsilon}$, де $\varepsilon > 0$ є згладжуючим множником порядку 10^{-8}, отримаємо ітераційну послідовність методу AdaGrad:

$$x_j^{k+1} = x_j^k - \bar{\rho}_k g_j^k, \; G_{k,j} = G_{k-1,j} + (g_j^k)^2, \; g^k = \nabla_x f(x^k) \; j = 1, \dots, n; \; k \in \mathbb{N} \tag{2.48}$$

Інтуїтивна інтерпретація правила оновлення (2.48) полягає в наступному: параметри, яким відповідають рідкісні, але інформативні ознаки, мають малу накопичену суму $G_{k,j}$ і, відповідно, отримують великий ефективний крок; натомість параметри з частими оновленнями швидко накопичують велику суму $G_{k,j}$, що автоматично зменшує їхній крок.

Для аналізу збіжності методу AdaGrad використовується формалізм стохастичної оптимізації з оцінкою різниці поточної ітерації та очікуваної точки інфімуму. Нехай на кожному кроці t алгоритм обирає точку $x_t \in X \subseteq \mathbb{R}^n$ та отримує субградієнт$g_t \in \partial f_t(x_t)$, тоді оцінка різниці визначається як [27]:

$$R(T) = \sum_{t=1}^{T} f_t(x_t) - \inf_{x \in X} \sum_{t=1}^{T} f_t(x) \tag{2.49}$$

Ключовою для доведення збіжності є наступна лема, що обмежує кумулятивний внесок адаптивно масштабованих градієнтів.

Лема 1 (Подвійна нерівність) [27] Нехай $g^t = \nabla f_t(x^t)$, і нехай $g_{1:t}$ та s_t визначені як в алгоритмі AdaGrad (2.48). Тоді:

$$\sum_{t=1}^{T} \langle g^t,\, \mathrm{diag}(s_t)^{-1} g^t \rangle \le 2 \sum_{j=1}^{n} \| g_{1:T,j} \|_2 \tag{2.50}$$

Комбінуючи лему 1 з оцінками дивергенції Брегмана $B_\psi(x, y) = \psi(x) - \psi(y) - \langle \psi(y), x - y \rangle$ для композитного дзеркального спуску, отримаємо основну теорему збіжності [19].

Теорема 8 (Збіжність AdaGrad) [27] Нехай послідовність $\{x^t\}$ визначена алгоритмом AdaGrad (2.48) з використанням оновлення композитного дзеркального спуску, і нехай $\max_t \|x^* - x^t\|_\infty \le D_\infty$. Тоді для довільного $x^* \in X$ при $\eta = D_\infty / \sqrt{2}$ виконується оцінка:

$$R_\varphi(T) \le \sqrt{2}\, D_\infty \sum_{j=1}^{n} \| g_{1:T,j} \|_2 = \sqrt{2n}\, D_\infty \sqrt{\inf_{s \succeq 0,\, \langle \mathbf{1}, s \rangle \le n} \sum_{t=1}^{T} \langle g^t, \mathrm{diag}(s)^{-1} g^t \rangle} \tag{2.51}$$

Ключовою перевагою оцінки (2.50) є наявність інфімуму під знаком кореня, що дозволяє автоматично обирати оптимальне масштабування, не гірше за будь-яку фіксовану діагональну проксимальну матрицю, обрану ретроспективно [27]. Зокрема, якщо область визначення $X = \{x : \|x\|_\infty \le 1\}$ є гіперкубом, то $D_\infty = 2$, і для розрідженого розподілу ознак з ймовірностями появи $p_j = \min\{1, c \cdot j^{-\alpha}\}$, $\alpha > 1$, оцінка різниці (2.48) методу AdaGrad становить $O(\max\{\log n, n^{1-\alpha/2}\}\sqrt{T})$, що може бути експоненціально меншим за оцінку $O(\sqrt{nT})$ неадаптивного градієнтного спуску.

Практичну ефективність AdaGrad на розріджених даних продемонстровано у задачі навчання векторних представлень слів GloVe [106]. У моделі GloVe мінімізується зважена функція найменших квадратів за матрицею спільної появи слів, що є розрідженою: більшість пар слів мають нульову або дуже малу частоту спільної появи. Застосування AdaGrad забезпечує великі кроки оновлення для рідкісних словесних пар, які несуть найбільшу семантичну інформацію, та малі

кроки для частих службових слів. Ця властивість робить метод AdaGrad доцільним для навчання мовних моделей DNN. Водночас метод AdaGrad має суттєвий недолік: оскільки кожний доданок $(g_j^t)^2$ є невід'ємним, накопичена сума $G_{k,j}$ монотонно зростає протягом навчання. Це призводить до поступового зменшення ефективного кроку $\bar{\rho}_k = \rho_k / \sqrt{G_k + \varepsilon}$, тобто $\lim\limits_{k\to\infty} \|\bar{\rho}_k\| = 0$. Як наслідок, на пізніх стадіях навчання алгоритм фактично припиняє оновлення параметрів, що важливо в задачах з нестаціонарними цільовими функціями. Наступні методи покликані подолати це обмеження.

Для вирішення проблеми неконтрольованого зменшення кроку в AdaGrad запропоновано метод RMSProp [121], що замінює кумулятивну суму квадратів градієнтів їхнім експоненціально зваженим ковзним середнім. Замість точного значення G_k використовується його стохастична апроксимація $\bar{G}_k \approx \mathbb{E}[G_k]$, контрольована коефіцієнтом усереднення $0 < \beta < 1$ [27]:

$$\begin{aligned} &\bar{G}_k = \beta \bar{G}_{k-1} + (1-\beta)\,\|g^k\|^2, \quad g^k = \nabla_x f(x^k), \\ &x_j^{k+1} = x_j^k - \frac{\rho_k}{\sqrt{\bar{G}_{k,j} + \varepsilon}} g_j^k, \qquad j = 1, \dots, n;\ k \in \mathbb{N}. \end{aligned} \tag{2.52}$$

На практиці, значення коефіцієнта $\beta = 0.9$ забезпечує ефективне «забування» старих градієнтів з шкалою приблизно $1/(1-\beta) = 10$ ітерацій. Завдяки цьому знаменник оновлення (2.52) залишається обмеженим навіть при великій кількості ітерацій, на відміну від монотонно зростаючого G_k у методі AdaGrad. Геометрична інтерпретація полягає в тому, що RMSProp адаптує крок не до всієї історії градієнтів, а лише до нещодавнього вікна, що робить метод придатним для нестаціонарних задач. Порівняно з AdaGrad, метод RMSProp зберігає перевагу покоординатної адаптації кроку навчання, але уникає передчасного затухання оновлень. Проте RMSProp не враховує інформацію про напрямок руху (момент першого порядку), а також не містить корекції зміщення початкових оцінок, що може призводити до нестабільності на ранніх етапах навчання при значеннях β близьких до одиниці.

Метод ADAM [56], поєднує переваги адаптивного кроку RMSProp з імпульсним прискоренням, одночасно розв'язуючи проблему зміщення початкових оцінок. Ідея полягає у підтримці експоненціально зважених ковзних середніх як для градієнта (момент першого порядку m_k), так і для квадрата градієнта (момент другого порядку v_k):

$$m_k = \beta_1\, m_{k-1} + (1-\beta_1)\, g^k,\ v_k = \beta_2\, v_{k-1} + (1-\beta_2)\,(g^k)^2 \tag{2.53}$$

де $g^k = \nabla_x f(x^k)$, а $0 < \beta_1 < 1$ та $0 < \beta_2 < 1$ є коефіцієнтами експоненціального згасання. Величина m_k представляє апроксимацію математичного сподівання градієнта $\mathbb{E}[g^k]$ і відіграє роль, аналогічну імпульсному доданку в методі «важкої кулі» (2.13), забезпечуючи згладжування осциляцій. Величина v_k апроксимує незцентровану дисперсію $\mathbb{E}[\|g^k\|^2]$ і масштабує крок навчання для кожного параметра.

Оскільки оцінки m_k та v_k ініціалізуються нульовими векторами, вони є зміщеними на початкових ітераціях [56]. Виведемо коригуючи множники для моменту другого порядку; для першого порядку обчислення є аналогічними. Розкриваючи рекурентне співвідношення (2.53) для v_k, отримаємо:

$$v_t = (1-\beta_2) \sum_{i=1}^{t} \beta_2^{t-i} \cdot (g^i)^2 \tag{2.54}$$

Обчислимо математичне сподівання обох частин. Припускаючи, що другий момент $\mathbb{E}[(g^i)^2]$ є стаціонарним (тобто однаковим для всіх i), отримуємо:

$$\begin{aligned} \mathbb{E}[v_t] &= \mathbb{E}\left[(1-\beta_2)\textstyle\sum_{i=1}^{t} \beta_2^{t-i} \cdot (g^i)^2\right] \\ &= \mathbb{E}[(g^t)^2] \cdot (1-\beta_2)\textstyle\sum_{i=1}^{t} \beta_2^{t-i} + \zeta \\ &= \mathbb{E}[(g^t)^2] \cdot (1-\beta_2^t) + \zeta, \end{aligned} \tag{2.55}$$

де $\zeta = 0$ у випадку стаціонарності; в іншому випадку ζ залишається малим за умови належного вибору β_2. Таким чином, $\mathbb{E}[v_t]$ відрізняється від $\mathbb{E}[(g^t)^2]$ на множник $(1-\beta_2^t)$, що свідчить про зміщення. Повторюючи аналогічні обчислення для першого моменту, отримуємо:

$$\mathbb{E}[m_t] = \mathbb{E}[g^t] \cdot (1-\beta_1^t) + \zeta' \tag{2.56}$$

Компенсуючи зміщення діленням на відповідні множники, отримаємо виправлені оцінки моментів першого та другого порядку:

$$\bar{m}^k = \frac{m^k}{1-\beta_1^k}, \qquad \bar{v}^k = \frac{v^k}{1-\beta_2^k} \tag{2.57}$$

Величини $\bar{m}^k$ та $\bar{v}^k$ є поточними незміщеними статистичними оцінками градієнтів $\nabla f(x^k)$ та квадратів норм $\|\nabla f(x^k)\|^2$ відповідно. Використовуючи ці виправлені оцінки, отримуємо ітераційну послідовність методу ADAM:

$$x^{k+1} = x^k - \frac{\rho_k}{\sqrt{\bar{v}^k + \varepsilon}}\,\bar{m}^k, \quad x^0 \in \mathbb{R}^n, \quad k \in \mathbb{N} \tag{2.58}$$

Рекомендовані значення гіперпараметрів: $\rho = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ та $\varepsilon = 10^{-8}$ [56]. Важливою властивістю правила оновлення (2.58) є інваріантність до масштабу градієнта: множення g^k на скаляр c масштабує $\bar{m}^k$ з множником c і $\bar{v}^k$ з множником c^2, що взаємно компенсується у дробі $(c \cdot \bar{m}^k)/\sqrt{c^2 \cdot \bar{v}^k} = \bar{m}^k/\sqrt{\bar{v}^k}$. Ефективний крок $\Delta_k = \rho_k \cdot \bar{m}^k/\sqrt{\bar{v}^k}$ є обмеженим приблизно величиною ρ_k, що встановлює так звану «область довіри» навколо поточного значення параметрів. Для аналізу збіжності методу ADAM необхідні допоміжні оцінки на адаптивно масштабовані градієнтні доданки.

Лема 2 [56] Нехай $g^t = \nabla f_t(x^t)$ з обмеженнями $\|g^t\|_2 \le G$ та $\|g^t\|_\infty \le G_\infty$, і нехай $\gamma = \beta_1^2/\sqrt{\beta_2} < 1$. Тоді:

$$\sum_{t=1}^{T} \frac{\hat{m}_{t,j}^2}{\sqrt{t\,\hat{v}_{t,j}}} \le \frac{2G_\infty}{(1-\gamma)^2\sqrt{1-\beta_2}}\,\|g_{1:T,j}\|_2 \tag{2.59}$$

Використовуючи отриману лему 2, можемо вивести оцінку збіжності для методу ADAM за допомогою наступної теореми:

Теорема 9 (Збіжність ADAM) [56] Нехай функції f_t мають обмежені градієнти: $\|\nabla f_t(x)\|_\infty \le G_\infty$ для всіх $x \in \mathbb{R}^n$, і відстань між послідовними точками обмежена: $\|x^s - x^r\|_\infty \le D_\infty$ для довільних $s, r \in \{1, \dots, T\}$. Нехай $\beta_1, \beta_2 \in [0, 1)$ задовольняють умову $\beta_1^2/\sqrt{\beta_2} < 1$, крок $\rho_t = \rho/\sqrt{t}$, а коефіцієнт першого моменту $\beta_{1,t} = \beta_1 \lambda^{t-1}$, $\lambda \in (0, 1)$. Тоді для методу ADAM виконується оцінка (2.49):

$$\begin{array}{ll} R(T) \le \frac{D^2}{2\rho(1-\beta_1)} \sum_{j=1}^{n} \sqrt{T\hat{v}_{T,j}} & + \\ \frac{\rho(1+\beta_1)G_\infty}{(1-\beta_1)\sqrt{1-\beta_2}(1-\gamma)^2} \sum_{j=1}^{n} \|g_{1:T,j}\|_2 & + \sum_{j=1}^{n} \frac{D_\infty^2 G_\infty \sqrt{1-\beta_2}}{2\rho(1-\beta_1)(1-\lambda)^2}. \end{array} \tag{2.60}$$

Порівняно з RMSProp, метод ADAM має дві ключові переваги. По-перше, наявність оцінки першого моменту $\bar{m}^k$ виконує роль імпульсного доданку, що згладжує осциляції градієнтного спуску та прискорює збіжність на плато цільової функції, аналогічно до методу «важкої кулі» (2.13). По-друге, корекція зміщення (2.57) запобігає надмірно великим крокам на початкових ітераціях, що, як показано емпірично [56], є важливим при значеннях β_2 близьких до одиниці – саме тих, які необхідні для стабільної роботи на розріджених градієнтах. На рисунку 2.2

показана траєкторія збіжності методів (2.40) та (2.58) при знаходженні мінімуму функції Розенброка.

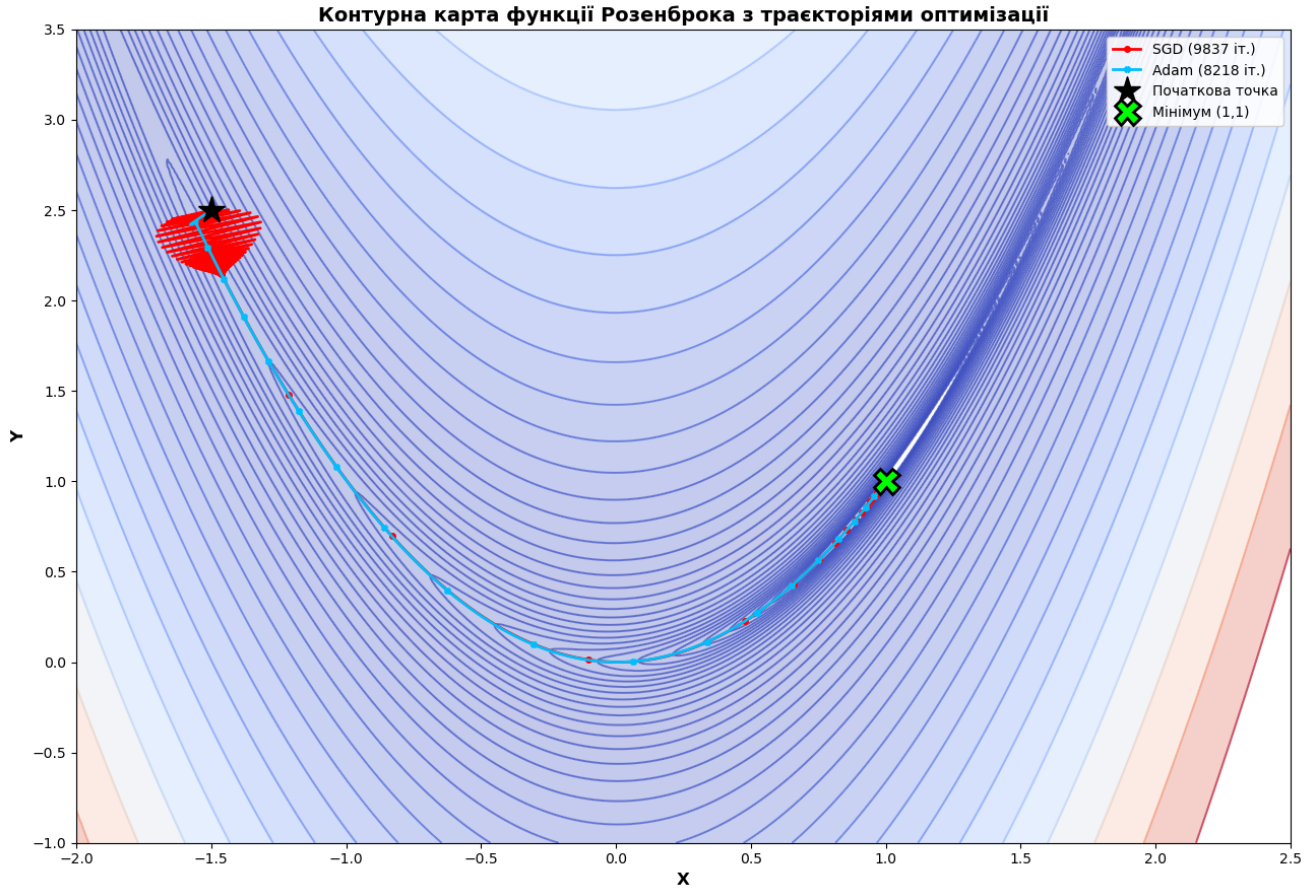

Рис. 2.2: Порівняння швидкості збіжності методу градієнтного спуску та методу «важкої кулі» при знаходженні мінімуму функції Розенброка $f(x,y) = (1-x)^2 + 100(y-x^2)^2$ при заданих параметрах $\rho = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ та початковій точці $(-1.5, 2.5)$.

Завдяки контролю параметрів адаптивної оцінки моментів ми отримаємо кращу швидкість збіжності. Практична ефективність методу ADAM яскраво проявилися у навчанні архітектури Transformer [13,48,106,123]. Модель Transformer, що лежить в основі сучасних великих мовних моделей, містить мільярди параметрів з суттєво різними масштабами градієнтів у різних шарах, як наприклад механізм Attention. Підсумовуючи, еволюція адаптивних методів від AdaGrad (2.48) через RMSProp (2.52) до ADAM (2.53) відображає послідовне вирішення ключових проблем градієнтної оптимізації:

1. AdaGrad вводить покоординатну адаптацію кроку для ефективної роботи з розрідженими даними;
2. RMSProp замінює кумулятивне накопичення ковзним середнім, запобігаючи передчасному затуханню кроку;
3. ADAM додає імпульсне прискорення та корекцію зміщення, забезпечуючи стабільну збіжність в широкому класі задач.

Кожний з цих методів зберігає обчислювальну складність $O(n)$ на ітерацію та потребує лише $O(n)$ додаткової пам'яті, що робить їх практично придатними для задач з мільярдами параметрів.

2.4 Висновки до розділу

Отже, у даному розділі було проведено огляд методів градієнтної оптимізації та встановлено основні теоретичні результати щодо їхньої збіжності. Аналітично виведено рекурентну формулу класичного методу градієнтного спуску (2.6) з означення градієнта та теореми Тейлора [32,133], встановлено умови збіжності при різних припущеннях на цільову функцію [82,124] та детально проаналізовано прискорені модифікації, а саме метод «важкої кулі» Поляка [99] та прискорений метод Нестерова [81]. Розглянуто теоретичні основи стохастичної апроксимації, закладені у початкових роботах Роббінса–Монро [104] та Кіфера–Вольфовіца [55], та їхній безпосередній зв'язок із методом стохастичного градієнтного спуску [32]. Проаналізовано послідовну еволюцію адаптивних методів оптимізації – AdaGrad [27], RMSProp [121] та ADAM [56], – що реалізують покоординатну адаптацію кроку навчання на основі накопиченої градієнтної інформації. Прискорений метод Нестерова досягає оптимальної для класу методів першого порядку швидкості збіжності $O(1/n^2)$ на опуклих функціях з ліпшицевим градієнтом [81,124], що є квадратичним покращенням порівняно зі швидкістю $O(1/n)$ класичного градієнтного спуску. Для сильно опуклих функцій методи Поляка та Нестерова забезпечують суттєве прискорення: їхня константа збіжності залежить від $\sqrt{M/m}$ замість M/m, що є важливим для погано обумовлених задач із великим числом обумовленості $\kappa = M/m$ [82,99]. Узагальнюючи властивості розглянутих методів, наведемо порівняльну характеристику їх збіжності в таблиці 2.1, де m та M –

параметри сильної опуклості та ліпшицевості градієнта відповідно, n – номер ітерації.

Табл. 2.1: Порівняння асимптотичної швидкості збіжності методу градієнтного спуску та його модифікацій

	Градієнтний спуск	**Метод Поляка**	**Метод Нестерова**
Опукла функція	$O(1/n)$	$O(1/n)$	$O(1/n^2)$
Сильно опукла функція	$O((1-m/M)^n)$	$O((1-\sqrt{m/M})^n)$	$O((1-\sqrt{m/M})^n)$

Незважаючи на такі переваги розглянутих методів як теоретичні гарантії збіжності та можливість масштабування до моделей DL з мільйонами параметрів [123], дані методи потребують щоб цільова функція була диференційованою, щоб гарантувати існування напрямку спуску до точки мінімуму. На практиці, в задачах оптимізації в основному зустрічаються негладкі цільові функції, де значення градієнта ∇f є невизначеним для певних точок, наприклад функції активації штучних нейронів ReLU [63]. Такий клас задач також включає в себе запропонований альтернативний метод векторної квантизації, який буде розглянутий в наступних розділах. Отже, має місце узагальнити розглянуті методи на негладкий випадок, використовуючи субдиференціал як узагальнений випадок градієнта ∇f.

РОЗДІЛ 3. УЗАГАЛЬНЕНІ МЕТОДИ ДЛЯ ОПТИМІЗАЦІЇ НЕГЛАДКИХ ЦІЛЬОВИХ ФУНКЦІЙ

У цьому розділі розглядаються основні відомості елементів негладкого аналізу, які дозволяють розширити методи стохастичної оптимізації для узагальнено-диференційовних функцій. Також приділено увагу тому як відповідність цільової функції властивостям Ліпшиця дозволить знайти точки екстремуму довільної задачі негладкої оптимізації.

Класична теорія оптимізації градієнтними методами передбачає умови, коли цільова функція є всюди диференційованою. Проте на практиці існує багато задач з функцією, яка має точки, де класичний градієнт не існує. Яскравими прикладами є задачі оптимізації моделей DL з негладкими функціями активації ReLU [63,123], комбінаторної оптимізації [26], кластеризації [103], реконструкції зображень [49] тощо. Це підкреслює потребу в розширенні поняття диференційовності та введенні додаткових необхідних умов існування екстремуму в негладких задачах.

Метод SQC має негладку цільову функцію, отже застосування класичних методів градієнтного спуску з попереднього розділу немає сенсу. Для можливості використання методів для даної задачі необхідно ввести поняття узагальненого градієнтного методу. Також в даному розділі буде розглянуто запропонований скінченно-різницевий метод для оптимізації негладких функції з використанням процедури згладжування за Стєкловим. Дослідження, представлені в даному розділі, були проілюстровані в публікації [93] та апробовані на конференції [85].

3.1 Оптимізація Ліпшицевих функцій

Функція $f : E_n \to E_1$ називається узагальнено-диференційованою в $x \in E_n$, якщо в деякому околі цієї точки існує багатозначне відображення у $\mapsto G_f(y)$, де $G_f(y)$ є непорожніми обмеженими опуклими замкненими множинами, і виконується розклад [39,88]:

$$f(y) = f(x) + \langle g, y - x \rangle + o(x, y, g), \quad g \in G_f(y) \tag{3.1}$$

де $\langle\cdot,\cdot\rangle$ позначає скалярний добуток, а функція залишку $o(x,y,g)$ задовольняє умову:

$$\lim_{\varepsilon\to 0}\sup_{\|y-x\|\le\varepsilon,\ y\neq x}\sup_{g\in G_f(y)}\frac{|o(x,y,g)|}{\|y-x\|}=0 \tag{3.2}$$

Для будь-яких послідовностей $y^k\to x$ ($y^k\neq x$), $g^k\in G_f(y^k)$ виконується $\lim_{k\to\infty} o(x,y^k,g^k)/\|y^k-x\|=0$. Вектори $g\in G_f(y)$ називаються узагальненими градієнтами функції f у точці y [88]. Функція називається узагальнено-диференційованою в області, якщо вона є узагальнено-диференційованою у кожній точці цієї області. Це означення є узагальненням поняття неперервної диференційовності: для опуклих функцій воно збігається з поняттям субградієнта, а для неопуклих функцій дозволяє визначити напрямок спуску навіть у точках, де класичний градієнт не існує. Для можливості пошуку точки екстремуму довільної задачі оптимізації необхідно, щоб цільова функція мала властивість ліпшицевості, яка забезпечує локальну обмеженість приростів функції. Функція $f: E_n\to E_1$ називається локально Ліпшицевою, якщо для будь-якої компактної множини $K\subset E_n$ існує константа L_K така, що [33,39]:

$$|f(x')-f(x'')|\le L_K\|x'-x''\|,\quad \forall\, x',x''\in K \tag{3.3}$$

Зв'язок між узагальненою диференційовністю та ліпшицевістю встановлюється наступним твердженням.

Теорема 10 (Ліпшицевість узагальнено-диференційовних функцій) [39, 88] Нехай $f: E_n\to E_1$ – узагальнено-диференційовна функція. Тоді f є локально Ліпшицевою, а отже, неперервною.

Нехай K – компактна множина в E_n, $\operatorname{co} K$ – її опукла оболонка. Оскільки відображення G_f є напів-неперервним зверху і його значення $G_f(x)$ є компактними множинами, воно є локально обмеженим; тому існує число $\Gamma_K=\sup\{\|g\|\mid g\in G_f(y),\ y\in\operatorname{co} K\}<+\infty$. Для $y,x\in K$ з розкладу (4.1) маємо [33]:

$$f(y)-f(x)=\left\langle g,\frac{y-x}{\|y-x\|}\right\rangle\|y-x\|+o(x,y,g),\quad g\in G_f(y) \tag{3.4}$$

Згідно з умовою (3.2), для будь-якого $\varepsilon>0$ існує $\delta(x,\varepsilon)>0$ такий, що $|o(x,y,g)|/\|y-x\|\le\varepsilon$ при $\|y-x\|\le\delta(x,\varepsilon)$, $g\in G_f(y)$. Звідси для кожного $x\in\operatorname{co} K$ у відповідному δ-околі виконується $|f(y)-f(x)|\le(\Gamma_K+\varepsilon)\|y-x\|$. Нехай тепер $x',x''\in K$; розглянемо відрізок $M=\{x'+t(x''-x'):t\in[0,1]\}\subset\operatorname{co} K$. Зазначені δ-

околи точок $x \in M$ покривають M; за компактністю існує скінченне покриття з центрами $x^1 = x', x^2, \dots, x^k = x''$. Для сусідніх точок x^j та x^{j+1} маємо $|f(x^j) - f(x^{j+1})| \le (\Gamma_K + \varepsilon)\|x^j - x^{j+1}\|$, звідки [10]:

$$|f(x') - f(x'')| \le \sum_{j=1}^{k-1} |f(x^j) - f(x^{j+1})| \le (\Gamma_K + \varepsilon)\|x' - x''\| \tag{3.5}$$

Оскільки Γ_K є сталою, а $\varepsilon > 0$ довільним, це показує, що функція є локально Ліпшицевою. Важливим наслідком є те, що неперервно диференційовні функції належать до класу узагальнено-диференційовних, причому їхні градієнти можна обрати як узагальнені. Нехай функція $f(y)$ ($y \in E_n$) є неперервно диференційованою з градієнтом $\nabla f(y)$. Тоді f є узагальнено-диференційованою з $G_f(y) = \{\nabla f(y)\}$. Іншим важливим класом функцій, що охоплюється узагальненою диференційовністю, є слабо опуклі функції. Неперервна функція $f : E_n \to E_1$ називається слабо опуклою, якщо для кожного x існує непорожня множина субградієнтів $G(x)$ таких, що для всіх $g \in G(x)$ справедлива нерівність [39]:

$$f(y) \ge f(x) + \langle g, y - x\rangle + r(x, y) \tag{3.6}$$

де $r(x,y)/\|y - x\| \to 0$ при $\|y - x\| \to 0$ рівномірно, коли x, y належать замкненій обмеженій множині. Для звичайних опуклих функцій $r(x, y) = 0$. Множина $G(x)$ є непорожньою, обмеженою, опуклою і замкненою, а багатозначне відображення $G : x \to G(x)$ є напів-неперервним зверху.

Теорема 11 (Узагальнена диференційовність слабо опуклих функцій) [39] Опуклі та слабо опуклі функції є узагальнено-диференційовними, а їхні субградієнти можна обрати як узагальнені.

Необхідно встановити справедливість розкладу (3.1) з умовою (3.2). Введемо позначення $o(x, y, g) = f(y) - f(x) - \langle g, y - x\rangle$. З нерівності (4.6), записаної відносно точки y, маємо $f(x) \ge f(y) + \langle g(y), x - y\rangle + r(y, x)$. Поєднуючи обидва співвідношення, отримаємо [88]:

$$\langle g(x) - g(y), y - x\rangle + r(x, y) \le o(x, y, g(y)) \le -r(y, x) \tag{3.7}$$

де $g(x) \in G(x)$, $g(y) \in G(y)$. Обираючи $g(x)$ так, щоб $\|g(x) - g(y)\| = \rho(g(y), G(x)) = \inf\{\|g(y) - g\| : g \in G(x)\}$, маємо [10]:

$$-\rho(g(y),G(x))\|y-x\|+r(x,y)\le o(x,y,g(y))\le -r(y,x) \qquad (3.8)$$

В силу напівнеперервності зверху відображення G та означення слабо опуклої функції при $\|y-x\|\to 0$ виконуються граничні співвідношення $\rho(g(y),G(x))\to 0$ та $r(x,y)/\|y-x\|\to 0$, $r(y,x)/\|y-x\|\to 0$, рівномірно по $g(y)\in G(y)$. Звідси випливає виконання умови (3.2). На практиці часто виникає необхідність оптимізації функцій, що є суперпозицією декількох узагальнено-диференційовних функцій. Наступне твердження встановлює замкненість класу таких функцій відносно композиції, узагальнюючи класичне правило ланцюжка диференціювання.

Теорема 12 (Правило ланцюжка для узагальнено-диференційовних функцій) [39] Нехай $f_0(z)$ ($z\in E_m$), $f_i(y)$ ($y\in E_n$, $i=1,\dots,m$) – узагальнено-диференційовні функції. Тоді складена функція $f(y)=f_0(f_1(y),\dots,f_m(y))$ також є узагальнено-диференційованою і:

$$G_f(y)=\operatorname{co}\{g\in E_n : g=[g^1\cdots g^m]\,g^0,\ g^0\in G_{f_0}(z(y)),\ g^i\in G_{f_i}(y),\ i=1,\dots,m\} \qquad (3.9)$$

де $[g^1\cdots g^m]$ – матриця розміру $n\times m$ зі стовпцями $g^1,\dots,g^m$, $z(y)=(f_1(y),\dots,f_m(y))$.

Таким чином, клас узагальнено-диференційовних функцій утворює лінійний простір і є замкненим відносно суперпозиції, що є важливою властивістю при оптимізації задач зі складеними цільовими функціями.

Взагалі, клас локально ліпшицевих функцій є широким, оскільки він охоплює практично всі класи неперервних функцій, що вивчаються в сучасній теорії оптимізації. Разом з тим локально ліпшицеві функції зберігають низку властивостей гладких функцій, зокрема, існування часткових похідних майже всюди. З того, що функція задовольняє умові Ліпшиця (3.3), випливає, що вона є абсолютно неперервною по кожній змінній x_i ($i=1,\dots,n$) при фіксованих інших змінних. Скінченна функція $g(t)$ на відрізку $[a,b]$ називається абсолютно неперервною, якщо для будь-якого $\varepsilon>0$ існує таке $\delta>0$, що для будь-якої скінченної системи взаємно непересічних інтервалів $(a_1,b_1),\dots,(a_p,b_p)$ з умови $\sum_{i=1}^{p}(b_i-a_i)<\delta$ випливає $\sum_{i=1}^{p}(b_i-a_i)<\delta$ [39,88]. Абсолютно неперервна функція має скінченну похідну майже всюди на $[a,b]$, тобто за винятком множини міри нуль. Користуючись цим, у локально ліпшицевої функції $f(x)$ майже всюди по змінній

x_i існує часткова похідна $\partial f(x)/\partial x_i$. Згідно з теоремою Фубіні, $f(x)$ має звичайну часткову похідну $\partial f/\partial x_i$ ($i = 1, \dots, n$) майже всюди в E_n.

Визначення 2 [39, 88] Локально ліпшицева функція $f(x)$, $x \in E_n$, є диференційованою майже всюди, тобто градієнт $\nabla f(x)$ існує всюди, за винятком множини міри нуль.

З'ясування необхідних умов екстремуму є важливим із кількох причин: у багатьох випадках ці умови визначають структуру оптимізаційних методів, відіграють ключову роль при дослідженні збіжності алгоритмів, а також дозволяють уникнути побудови окремих критеріїв для кожного конкретного випадку. На відміну від опуклих функцій, необхідні умови екстремуму для узагальнено-диференційовних і ліпшицевих функцій характеризують лише локальні властивості: якщо в точці $x = x^*$ ці умови виконуються, то за відсутності додаткових припущень можна лише стверджувати, що в ній досягається локальний екстремум.

Теорема 13 (Умова Куна-Такера) [39, 88] Нехай дано задачу $f(x) \to \min$, $h(x) \le 0$, де $f(x)$, $h(x)$ – узагальнено-диференційовні функції. Визначимо багатозначне відображення $G : x \to G(x)$:

$$G(x) = \begin{cases} G_f(x), & h(x) < 0, \\ \operatorname{co}\{G_f(x) \cup G_h(x)\}, & h(x) = 0, \\ G_h(x), & h(x) > 0. \end{cases} \tag{3.10}$$

Нехай x^* – точка мінімуму $f(x)$ при обмеженні $h(x) \le 0$. Тоді $0 \in G(x^*)$. Якщо, крім того, виконано умову регулярності $\inf\{\|g\| : g \in G_h(x^*),\ h(x^*) = 0\} = \gamma > 0$, то існують такі $\lambda^* \ge 0$, $g_f^* \in G_f(x^*)$, $g_h^* \in G_h(x^*)$, що:

$$g_f^* + \lambda^* g_h^* = 0, \lambda^* h(x^*) = 0 \tag{3.11}$$

Для локально ліпшицевих функцій необхідні умови екстремуму формулюються через узагальнений градієнт Кларка $\partial f(x)$.

Теорема 14 (Необхідна умова мінімуму) [39, 88] Нехай $f(x)$ – локально ліпшицева функція на множині $x \in E_n$. Для того щоб $f(x)$ досягала в точці $x = x^*$ локального мінімуму, необхідно, щоб $0 \in \partial f(x^*)$.

Перевагою репрезентації через градієнт Кларка $\partial f(x)$ полягає у можливості узагальнити пошук стаціонарних точок для таких негладких функцій як $\min f(x)$ або $\max f(x)$.

Теорема 15 (Субградієнт функції мінімуму) [39, 88] Нехай $f(x) = f(x_1, \dots, x_n)$ – локально ліпшицева функція на множині $x \in E_n$. Тоді існує єдиний субградієнт мінімуму функції:

$$\partial(\min_{1 \le k \le n} f(x_1, \dots, x_n)) = \partial f(x_{k^*}), k^* = \arg\min_{1 \le k \le n} \{f(x_k)\}_n \tag{3.12}$$

3.2 Узагальнений градієнтний метод

Встановивши необхідні умови екстремуму для узагальнено-диференційовних та ліпшицевих функцій, перейдемо до побудови алгоритмічного методу пошуку стаціонарних точок. Класичний градієнтний спуск, призначений для гладких задач, не може бути безпосередньо застосований до функцій, що не мають похідних за напрямками. Проте поняття узагальненої диференційовності дозволяє розширити застосування градієнтної ітераційної послідовності на клас недиференційованих функцій. Зазначимо, що дослідження збіжності чисельних методів неопуклої нелінійної оптимізації є помітно складнішим, ніж у випадку опуклих або гладких задач, оскільки в таких задачах часто відсутня монотонна зміна цільової функції вздовж породжених послідовностей, що унеможливлює пряме застосування стандартних релаксаційних аргументів [39,88].

Розглянемо задачу безумовної мінімізації узагальнено-диференційованої функції:

$$f(x) \to \min, \quad x \in E_n \tag{3.13}$$

де $f : E_n \to E_1$ є узагальнено-диференційованою функцією у сенсі означення (3.1). Оскільки узагальнено-диференційовні функції, взагалі кажучи, не мають похідних за напрямками, структура задачі суттєво відрізняється від класичного гладкого випадку. Множину стаціонарних точок задачі (3.13) визначимо як [88]:

$$X^* = \{x \in E_n \mid 0 \in G_f(x)\} \tag{3.14}$$

де $G_f(x)$ – множина узагальнених градієнтів функції f у точці x. Множина X^* є замкненою внаслідок напівнеперервності зверху псевдоградієнтного відображення G_f [39]. Згідно з теоремою 14, усі точки локального мінімуму функції $f(x)$ містяться у множині X^*. Введемо також множину стаціонарних значень $F^* = \{f(x) \mid x \in X^*\}$.

Метод узагальненого градієнта будується як пряме розширення класичного градієнтного спуску на випадок узагальнено-диференційовних функцій шляхом заміни градієнта довільним елементом множини узагальнених градієнтів [39,88]:

$$x^{k+1} = x^k - \rho_k g^k, \quad g^k \in G_f(x^k), \quad k = 0, 1, ... \tag{3.15}$$

де послідовність кроків $\{\rho_k\}_{k=0}^{\infty}$ задовольняє стандартні умови:

$$0 \le \rho_k \le R, \quad \lim_{k\to\infty} \rho_k = 0, \quad \sum_{k=0}^{\infty} \rho_k = +\infty \tag{3.16}$$

Перша умова забезпечує обмеженість кроків, друга гарантує їх прямування до нуля, а третя виключає передчасну зупинку алгоритму. Зауважимо, що в неопуклому випадку вибір конкретного узагальненого градієнта $g^k \in G_f(x^k)$ на кожній ітерації є довільним, що відрізняє цей метод від субградієнтних методів опуклої оптимізації, де субградієнт визначається однозначно із субдиференціалу.

Для дослідження збіжності методу (3.15) необхідно спочатку встановити обмеженість породжуваних послідовностей. Наступна лема показує, що за достатньо малих кроків траєкторії методу не виходять з обмеженої області.

Лема 3 (Обмеженість послідовностей) [88] Нехай існує $c > f(x^*)$, що не належить F^*. Тоді для будь-якого $d > c$ знайдеться таке $R > 0$, що для будь-якої послідовності $\{\rho_k\}_{k=0}^{\infty}$, яка задовольняє умовам (4.16) із $\sup \rho_k \le R$ та $f(x^0) \le c$, послідовність $\{x^k\}$, породжена рекурентним співвідношенням (4.15), не виходить з обмеженої множини $\{x \mid f(x) \le d\}$.

Зазначимо, що потрібне в лемі число c існує, зокрема, якщо множина F^* обмежена або якщо F^* не містить інтервалів. На практиці величина $R = \sup \rho_k$ підбирається наступним чином: алгоритм запускається з точки x^0 з деякою послідовністю $\{\rho_k\}$ та $R_0 = \sup \rho_k$. Якщо на деякому кроці виявляється, що $f(x^k) > d$, процес переривається, обирається нова послідовність $\{\rho'_k\}$ із меншим $R_1 = \sup \rho'_k$, і алгоритм перезапускається з початкової точки x^0. Оскільки $R_s \to 0$ при $s \to \infty$, рано чи пізно буде $R_s < R$, і відповідна траєкторія залишиться в обмеженій області [39].

Центральним елементом аналізу збіжності є встановлення локальної мінімізуючої властивості методу: запуск із нестаціонарної точки з достатньо малим кроком гарантує знаходження меншого значення функції в малому околі точки старту. Для доведення цієї властивості використовується така геометрична лема.

Лема 4 (Геометрична властивість опуклої множини) [88] Нехай P – опукла множина в E_n, причому $0 < \gamma \le \|p\| \le \Gamma < +\infty$ для всіх $p \in P$. Тоді для будь-якого набору векторів $\{p^r \in P \mid r = k, \dots, m\}$ і будь-якого набору чисел $\{\rho_r \ge 0 \mid r = k, \dots, m\}$ таких, що $\sum_{r=k}^{m} \rho_r \ge \sigma > 0$ та

$$\sup_{k \le r \le m} \rho_r \le R = \frac{\gamma\sigma}{6\Gamma} \tag{3.17}$$

знайдеться індекс $l \in (k, m]$, для якого

$$\left\langle p^l, \frac{\sum_{r=k}^{l-1} \rho_r p^r}{\sum_{r=k}^{l-1} \rho_r} \right\rangle > \frac{\gamma^2}{4}, \quad \sum_{r=k}^{l-1} \rho_r \ge \frac{\gamma\sigma}{3\Gamma} \tag{3.18}$$

Встановивши геометричну властивість, сформулюємо ключову лему про локальну мінімізацію, яка встановлює зв'язок між нестаціонарністю точки та здатністю методу зменшити значення цільової функції.

Лема 5 (Локальна мінімізуюча властивість) [39, 88] Нехай послідовність початкових точок $\{x^s\}_{s=0}^{\infty}$ збігається до $x = \lim_{s\to\infty} x^s$ і $0 \notin G_f(x)$. Нехай для кожного s послідовність $\{x_k^s\}_{k \ge k_s}$ побудована за правилом (4.15) з початковою точкою $x_{k_s}^s = x^s$ та кроками ρ_k^s $(k \ge k_s)$ до деякого моменту n_s. Позначимо $\bar\rho_s = \sup_{k \ge k_s} \rho_k^s$ та $\sigma_s = \sum_{k=k_s}^{n_s-1} \rho_k^s$. Якщо $\sigma_s \ge \sigma > 0$ і $\lim\limits_{s\to\infty} \bar\rho_s = 0$, то існує таке $\bar\varepsilon > 0$, що для будь-якого $\varepsilon \in (0, \bar\varepsilon]$ знайдуться індекси l_s такі, що для достатньо великих s:

$$\|x_k^s - x\| \le \varepsilon \quad k \in [k_s, l_s], \quad f(x) = \lim_{s\to\infty} f(x^s) > \overline{\lim_{s\to\infty}} f(x_{l_s}^s) \tag{3.19}$$

Лема 5 використовується для доведення основного результату про збіжність методу узагальненого градієнта. Вона стверджує, що в околі будь-якої нестаціонарної точки алгоритм здатний зменшити значення цільової функції, що є необхідною умовою для прогресу оптимізаційного процесу. Тепер сформулюємо та доведемо основну теорему про збіжність.

Теорема 16 (Збіжність методу узагальненого градієнта) [39, 88] Нехай $f : E_n \to E_1$ – узагальнено-диференційовна функція, і послідовність $\{x^k\}_{k=0}^{\infty}$, породжена методом (4.15) з умовами (4.16), обмежена. Тоді:

1. мінімальні за значенням $f(x)$ граничні точки послідовності $\{x^k\}_{k=0}^{\infty}$ належать множині стаціонарних точок X^*;
2. множина всіх граничних точок числової послідовності $\{f(x^k)\}_{k=0}^{\infty}$ утворює інтервал, що лежить у множині F^*;

3. якщо множина F^* не містить інтервалів зокрема, є скінченною або зліченною, то числова послідовність $\{f(x^k)\}_{k=0}^{\infty}$ має границю, а всі граничні точки $\{x^k\}_{k=0}^{\infty}$ належать зв'язним компонентам множини X^*.

Теорема 16 встановлює збіжність методу узагальненого градієнта для неопуклих задач у тому сенсі, що мінімальні за значенням цільової функції граничні точки породженої послідовності є стаціонарними, тобто задовольняють необхідну умову мінімуму $0 \in G_f(x^*)$ з теореми 14. Зауважимо, що у загальному неопуклому випадку не гарантується збіжність усієї послідовності до однієї точки; проте якщо множина стаціонарних значень F^* дискретна, числова послідовність значень функції має границю, що істотно спрощує аналіз.

Метод (3.15) має головним чином теоретичне значення через повільну збіжність. Одним із шляхів його покращення є нормування узагальнених градієнтів. Позначимо нормований вектор:

$$\bar{g} = \begin{cases} \dfrac{g}{\|g\|}, & g \neq 0, \\ 0, & g = 0, \end{cases} \tag{3.20}$$

і розглянемо метод нормованого узагальненого градієнта $x^{k+1} = x^k - \rho_k \bar{g}^k$, $\bar{g}^k \in \overline{G_f(x^k)}$. Для цього методу локальна мінімізуюча властивість, встановлена лемою 5, залишається справедливою. Дійсно, подавши $x^{k+1} = x^k - \bar{\rho}_k g^k$ з $\bar{\rho}_k = \rho_k / \|g^k\|$ при $g^k \neq 0$, можна показати, що послідовність модифікованих кроків $\{\bar{\rho}_k\}$ задовольняє умовам леми 5 у малому околі нестаціонарної точки, оскільки $\rho_k/\Gamma_1 \leq \bar{\rho}_k \leq 2\rho_k/\eta_1$, де $\eta_1 = \rho(0, G_f(x)) > 0$ та $\Gamma_1 = \sup\{\|g\| \mid g \in G_f(y),\ \|y - x\| \leq \varepsilon_1\}$ [88]. Відповідно, теорема 16 повністю поширюється на метод нормованого узагальненого градієнта.

3.3 Згладжування за Стєкловим

В дослідженнях [8,28,82] було запропоновано метод згладжування функції, а саме усереднення на гіперкубах, застосований для дослідження та оптимізації негладких функцій Ліпшиця з використанням стохастичних скінченно-різницевих методів. Щоб гарантувати локальну збіжність методу оптимізації до стаціонарних точок задачі, параметр згладжування прямує до нуля, узгоджуючись з ітераційними кроками методу. Тут ми узагальнюємо задачу стохастичної оптимізації (1.30) за

допомогою методу згладжування усереднених функцій (також відомого як згладжування Стеклова [39,87,88]). У роботі [10] було запропоновано оцінку градієнта за допомогою згладжування на сфері, де ми оцінюємо очікуване значення за багатовимірним рівномірним розподілом на кулі $B_h(x)$ з радіусом h та об'ємом ν_n для кожного x, тобто $B_h(x) = \{y \in \mathbb{R}^n | ||y - x|| \leq h\}$:

$$f_\varepsilon(x) = \frac{1}{\nu_n} \int_{B_h(x)} f(y)dy \tag{3.21}$$

Неперервність Ліпшиця $f_\varepsilon(x)$ є диференційованою, її субдиференційовна множина дорівнює очікуваному значенню диференційованого стохастичного оракула [28]:

$$\nabla f_\varepsilon(x) = \frac{1}{\nu_n} \int_{B_h(x)} \partial f(y)dy \tag{3.22}$$

Оскільки за означенням функція $f(x_k)$ є неперервною за Ліпшицем, ми можемо виразити згладжену функцію $f_\varepsilon(x)$ за допомогою поверхневих інтегралів:

$$\begin{aligned} \nabla f_\varepsilon(x) &= \tfrac{1}{s_n} \textstyle\int_{S_h(x)} F(x+y)N(y)dS = \\ &= \tfrac{1}{h} \textstyle\int_{S_h(x)} (F(x+y) - F(x))N(y)dS = \\ &= \tfrac{1}{2h} \textstyle\int_{S_h(x)} (F(x+y) - F(x-y))N(y)dS \end{aligned} \tag{3.23}$$

де s_n – площа сферичної поверхні $S_h(x)$, а $N(y)$ – зовнішній нормальний вектор до поверхні $S_h(x)$ для певного елемента $N(y)$. Ми також можемо оцінити поверхневий інтеграл (3.23) за допомогою алгоритму Монте-Карло на множині K рівномірно розподілених стохастичних векторів $\bar{y}$ на одній сфері $S_1(0) = \{y \in \mathbb{R}^n : ||y|| = 1\}$. Тепер ми можемо подати згладжену функцію як скінченно-різницеву суму вздовж стохастичних векторів:

$$\begin{aligned} \nabla f_\varepsilon(x) &= \tfrac{n}{h} \mathbb{E}_{\bar{y}}[f(x + h\bar{y}) - f(x)]\bar{y} \\ &= \tfrac{n}{2h} \mathbb{E}_{\bar{y}}[f(x + h\bar{y}) - f(x - h\bar{y})]\bar{y} \\ &= \mathbb{E}_{\bar{y}}[||\tfrac{1}{2h}(f(x + h\bar{y}) - f(x - h\bar{y}))\bar{y}||^2] \leq L^2 \end{aligned} \tag{3.24}$$

Якщо цільову функцію f визначено на рівномірному розподілі ймовірностей $\mu(x)$ на просторі $\mathbb{R}^n$, який має властивості $\mu(x) \geq 0$ і $\int_{\mathbb{R}^n} \mu(x)dx = 1$, можна застосувати згладжену стохастичну апроксимацію (3.21) в околі $B_h(x) = \mathbb{R}^n$ і отримати згладжену апроксимацію $\nabla f_\varepsilon(x)$ зі змінною згладжування $h > 0$, яка також являє собою розмір інтервалу згладжування:

$$\begin{aligned} f_\varepsilon(x) &= \tfrac{1}{h^n} \textstyle\int_{\mathbb{R}^n} f(x+y)\mu(\frac{y}{h})dy \\ &= \textstyle\int_{\mathbb{R}^n} f(x+hz)\mu(z)dz \\ &= \tfrac{1}{h^n} \textstyle\int_{\mathbb{R}^n} f(y)\mu(\frac{y-x}{h})dy \end{aligned} \tag{3.25}$$

Зауважимо, що якщо функція є неперервною за Ліпшицем, то вона також є диференційованою (оскільки швидкість зміни функції за аргументом завжди скінченна), тому градієнт функції можна описати за допомогою $\nabla f_\varepsilon(x)$ підстановки виразу всередині інтеграла в (3.25):

$$\begin{array}{ll} \nabla f_\varepsilon(x) & = \frac{1}{h^n}\int_{\mathbb{R}^n} \partial f(x+y)\mu(\frac{y}{h})dy \\ & = \int_{\mathbb{R}^n} \partial f(x+hz)\mu(z)dz \end{array} \quad (3.26)$$

Використовуючи те саме наближення, що і в (3.22), (3.23), ми можемо замінити значення градієнта на згладжену функцію і отримати узагальнений алгоритм спуску за стохастичним градієнтом:

$$x_{k+1} = x_k - \rho_k \eta_k,\ \eta_k = \frac{1}{2h_k}(F(x_k + h_k\bar{y}) - F(x_k - h_k\bar{y}))\bar{y} \quad (3.27)$$

Ще однією важливою перевагою узагальненої форми градієнтного спуску порівняно з класичною формою (3.38) є те, що швидкість збіжності можна оцінювати на негладких функціях.

Лема 6 (Міра множини) [112] Для довільної неперервної L-ліпшицевої функції $f(x)$, якщо ξ є рівномірним розподілом на одиничній сфері $S_1(0)$ в n-вимірному евклідовому просторі, то

$$\sqrt{\mathbb{E}[(f(\xi) - \mathbb{E}[f(\xi)])^4]} \le C\frac{L^2}{n} \quad (3.28)$$

де C – деяка числова константа.

Теорема 17 (Швидкість збіжності кінцево-різницевого методу) [93] Нехай цільова функція $F(\cdot)$ є опуклою та ліпшицевою з деякою константою L (в евклідовій нормі) на опуклому компакті $X \subseteq \mathrm{R}^n$ та оцінкою $\|X\| = \sup_{x\in X}\|x\| \le D$, а послідовність $\{x_t\}$ будується рекурентно за наступним рівнянням:

$$x_{t+1} = \pi_X\left(x_t - \rho_t\eta_t\right), \quad x_1 \in X, \quad t = 1, 2, \dots, \quad (3.29)$$

а випадкові напрямки η_t в (3.29) обчислюються як:

$$\eta_t = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{2h_t}\left(F(x_t + h_t\tilde{y}_t^k) - F(x_t - h_t\tilde{y}_t^k)\right)\tilde{y}_t^k, \quad (3.30)$$

де $\{\tilde{y}_t^k\}_{k=1}^K$ – незалежні випадкові вектори, рівномірно розподілені на одиничній сфері з центром у початку координат. Тоді при постійному кроці $\rho = (D\sqrt{nK})/(L\sqrt{2T(C+K/n)})$ та постійному параметрі згладжування $h_t = h$ середня оцінка $\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t$ задовольняє вимоги:

$$\mathbb{E}F_h(\bar{x}_T) - \min_{x\in X} F_h(x) \le \frac{LD}{\sqrt{T}}\sqrt{\frac{2n}{K}}\left(C + \frac{K}{n}\right)^{1/2} \tag{3.31}$$

Для змінних кроків $\rho_t = (D\sqrt{nK})/(L\sqrt{2t(C+K/n)})$ та $\bar{x}_t = \sum_{\tau=1}^{t}\rho_t x_t / \sum_{\tau=1}^{t}\rho_t$ має місце оцінка:

$$\mathbb{E}F_h(\bar{x}_t) - \min_{x\in X} F_h(x) \le \frac{DL}{\sqrt{2}}\sqrt{\frac{n}{K}}\left(C + \frac{K}{n}\right)^{1/2}\frac{2+\ln t}{\sqrt{t}-1} \tag{3.32}$$

де C є деякою абсолютною константою з нерівності (3.28).

Доведення. Нехай x^* є довільною точкою в X. Згідно з властивістю опуклості X виконується наступна нерівність:

$$\begin{aligned}\|x^* - x_{t+1}\|^2 &= \|x^* - \pi_X(x_t - \rho_t\eta_t)\|^2 \le \|x^* - x_t + \rho_t\eta_t)\|^2 \\ &= \|x^* - x_t\|^2 + 2\rho_t\langle\eta_t, x^* - x_t\rangle + \rho_t^2\|\eta_t\|^2\end{aligned} \tag{3.33}$$

Візьмемо умовне математичне сподівання від обох частин цієї нерівності відносно σ-алгебри, породженої випадковою величиною $\{x_t\}$:

$$\mathbb{E}\left[\|x^* - x_{t+1}\|^2\middle| x_t\right] \le \|x^* - x_t\|^2 + \frac{2\rho_t}{n}\left\langle\nabla F_{h_t}(x_t), x^* - x_t\right\rangle + \rho_t^2\mathbb{E}\left[\|\eta_t\|^2\middle| x_t\right] \tag{3.34}$$

Внаслідок властивості опуклості функцій F та F_{h_t} маємо: $\left\langle\nabla F_{h_t}(x_t), x^* - x_t\right\rangle \le F_{h_t}(x^*) - F_{h_t}(x_t)$. В результаті отримуємо нерівність

$$\begin{aligned}2\rho_t\left(F_{h_t}(x_t) - F_{h_t}(x^*)\right) &\le -n\mathbb{E}\left[\|x^* - x_{t+1}\|^2\middle| x_t\right] \\ &\quad + n\|x^* - x_t\|^2 + \rho_t^2 n\mathbb{E}\left[\|\eta_t\|^2\middle| x_t\right]\end{aligned} \tag{3.35}$$

Візьмемо математичне сподівання обох частин цієї нерівності:

$$\begin{aligned}2\rho_t\left(\mathbb{E}F_{h_t}(x_t) - F_{h_t}(x^*)\right) &\le -n\mathbb{E}\|x^* - x_{t+1}\|^2 \\ &\quad + n\mathbb{E}\|x^* - x_t\|^2 + \rho_t^2 n\mathbb{E}\mathbb{E}\left[\|\eta_t\|^2\middle| x_t\right]\end{aligned} \tag{3.36}$$

Позначимо $\eta_t^k = \left(F(x_t + h_t\tilde{y}_t^k) - F(x_t - h_t\tilde{y}_t^k)\right)\tilde{y}_t^k/(2h_t)$ та відповідну оцінку як $\bar{\eta}_t = \mathbb{E}\left[\eta_t^k|x_t\right] = \mathbb{E}\left[\left(F(x_t + h_t\tilde{y}_t^1) - F(x_t - h_t\tilde{y}_t^1)\right)\tilde{y}_t^1/2h_t|x_t\right]$. Відповідно до рівняння (3.24) маємо, $\mathbb{E}\left[\eta_t^k|x_t\right] = \bar{\eta}_t = n^{-1}\nabla F_{h_t}(x_t)$ для всіх k. Для виразу $\mathbb{E}\left[\|\eta_t\|^2\middle| x_t\right]$ можемо отримати наступні оцінки:

$$\begin{aligned}\mathbb{E}\left[\|\eta_t\|^2\middle| x_t\right] &= \mathbb{E}\left[\left\|\tfrac{1}{K}\sum_{k=1}^{K}\eta_t^k\right\|^2|x_t\right] = \mathbb{E}\left[\left\|\tfrac{1}{K}\sum_{k=1}^{K}(\eta_t^k - \bar{\eta}_t) + \bar{\eta}_t\right\|^2|x_t\right] \\ &= \tfrac{2}{K^2}\sum_{k=1}^{K}\mathbb{E}\left[\left(\eta_t^k - \bar{\eta}_t\right)^2|x_t\right] + 2\mathbb{E}\left[\|\bar{\eta}_t\|^2|x_t\right] \\ &\le \tfrac{2}{K^2}\sum_{k=1}^{K}\left(\mathbb{E}\left[\|\eta_t^k\|^2|x_t\right] - \mathbb{E}\left[\|\bar{\eta}_t\|^2|x_t\right]\right) + 2\mathbb{E}\left[\|\bar{\eta}_t\|^2|x_t\right] \\ &\tfrac{2CL^2}{Kn} + \tfrac{2(1-1/K)}{n^2}\left\|\nabla F_{h_t}(x_t)\right\|^2 \le \tfrac{2CL^2}{Kn} + \tfrac{2L^2}{n^2}\left(1 - \tfrac{1}{K}\right)\end{aligned} \tag{3.37}$$

В результаті маємо нерівність:

$$\begin{aligned} D(x_t) &= 2\rho_t \left(\mathbb{E}F_{h_t}(x_t) - F_{h_t}(x^*)\right) \\ &\le n\mathbb{E}\|x^* - x_t\|^2 - n\mathbb{E}\|x^* - x_{t+1}\|^2 + \rho_t^2 n\mathbb{E}\mathbb{E}\left[\|\eta_t\|^2 \middle| x_t\right] \\ &\le n\mathbb{E}\|x^* - x_t\|^2 - n\mathbb{E}\|x^* - x_{t+1}\|^2 + \frac{2\rho_t^2 L^2}{K}\left(C + \frac{K}{n}\right) \end{aligned} \tag{3.38}$$

Тепер обчислимо суми нерівності по t від 1 до T:

$$2\sum_{t=1}^{T}\rho_t\left(\mathbb{E}F_{h_t}(x_t) - F_{h_t}(x^*)\right) \le n\mathbb{E}\|x^* - x_1\|^2 + \frac{2L^2}{K}\left(C + \frac{K}{n}\right)\sum_{t=1}^{T}\rho_t^2 \tag{3.39}$$

Спочатку розглянемо випадок фіксованого параметра згладжування $h_t = h$ для всіх t та нехай x^* задовольняє $F_h(x^*) = \min_{x\in X} F_h(x)$. Розділимо обидві частини вищенаведеної нерівності на $2\sum_{t=1}^{T}\rho_t$. Для середньої точки $\bar{x}_T = \sum_{t=1}^{T}\rho_t x_t / \sum_{t=1}^{T}\rho_t$ через властивість опуклості функції F_h маємо:

$$\begin{aligned} \mathbb{E}F_h(\bar{x}_t) - \min_{x\in X} F_h(x) &\le \frac{\sum_{t=1}^{T}\rho_t(\mathbb{E}F_h(x_t) - \min_{x\in X} F_h(x))}{\sum_{t=1}^{T}\rho_t} \\ &\le \frac{n\mathbb{E}\|x^* - x_1\|^2}{2\sum_{t=1}^{T}\rho_t} + \frac{L^2}{K}\left(C + \frac{K}{n}\right)\frac{\sum_{t=1}^{T}\rho_t^2}{\sum_{t=1}^{T}\rho_t} \end{aligned} \tag{3.40}$$

Для постійної величини кроку $\rho_t \equiv \rho$ та $\bar{x}_T = T^{-1}\sum_{t=1}^{T}\rho_t$, маємо

$$\mathbb{E}F_h(\bar{x}_T) - \min_{x\in X} F_h(x) \le \frac{nD^2}{2T\rho} + \frac{L^2\rho}{K}\left(C + \frac{K}{n}\right) \tag{3.41}$$

Для мінімізації правої частини (3.40) позначимо $\rho = (D\sqrt{nK})/(L\sqrt{2T(C + K/n)})$, тоді

$$\mathbb{E}F_h(\bar{x}_T) - \min_{x\in X} F_h(x) \le \frac{LD}{\sqrt{T}}\sqrt{\frac{2n}{K}}\left(C + \frac{K}{n}\right)^{1/2} \tag{3.42}$$

Для змінної величини кроку $\rho_t = (D\sqrt{nK})/(L\sqrt{2t(C + K/n)})$ та $\bar{x}_t = \sum_{\tau=1}^{t}\rho_\tau x_\tau / \sum_{\tau=1}^{t}\rho_\tau$, з (3.40) отримуємо

$$\begin{aligned} \mathbb{E}F_h(\bar{x}_t) - \min_{x\in X} F_h(x) &= \frac{DL}{\sqrt{2}}\sqrt{\frac{n}{K}}\left(C + \frac{K}{n}\right)^{1/2}\frac{\left(1 + \sum_{\tau=1}^{t}\tau^{-1}\right)}{\sum_{\tau=1}^{t}\tau^{-1/2}} \\ &\le \frac{DL}{\sqrt{2}}\sqrt{\frac{n}{K}}\left(C + \frac{K}{n}\right)^{1/2}\frac{2 + \ln t}{\sqrt{t} - 1} \end{aligned} \tag{3.43}$$

Тепер розглянемо випадок змінного параметра згладжування h_t та нехай x^* задовольняє умові $F(x^*) = \min_{x\in X} F(x)$. Оскільки $|F(x) - F_{h_t}(x)| \le Lh_t$ для $x \in X$, то $|F_{h_t}(x_t) - F(x_t)| \le Lh_t$, $|F_{h_t}(x^*) - F(x^*)| \le Lh_t$, та має місце наступна оцінка:

$$\begin{aligned} 2\sum_{t=1}^{T}\rho_t\left(\mathbb{E}F(x_t)\right. &- \left.\min_{x\in X} F(x)\right) = 2\sum_{t=1}^{T}\rho_t\left(\mathbb{E}F(x_t) - F(x^*)\right) \\ &\le 2\sum_{t=1}^{T}\rho_t\left(\mathbb{E}F_{h_t}(x_t) - F_{h_t}(x^*)\right) + 4\sum_{t=1}^{T}L\rho_t h_t \\ &\le nD^2 + \frac{2L^2}{K}\left(C + \frac{K}{n}\right)\sum_{t=1}^{T}\rho_t^2 + 4\sum_{t=1}^{T}L\rho_t h_t \end{aligned} \tag{3.44}$$

Поділивши дану нерівність на $\sum_{t=1}^{t}\rho_\tau$ для $\bar{x}_t = \sum_{\tau=1}^{t}\rho_\tau x_\tau / \sum_{\tau=1}^{t}\rho_\tau$ та $\rho_t = (D\sqrt{nK})/(L\sqrt{2t(C + K/n)})$, $h_t = L\rho_t/K$, отримаємо:

$$\begin{aligned}\mathbb{E}F(\bar{x}_t) - \min_{x\in X} F(x) &\leq \frac{nD^2}{2\sum_{\tau=1}^t \rho_\tau} + \frac{L^2}{K}\left(C + \frac{K}{n}\right)\frac{\sum_{\tau=1}^t \rho_\tau^2}{\sum_{\tau=1}^t \rho_\tau} + 2L\frac{\sum_{\tau=1}^t \rho_\tau h_\tau}{\sum_{\tau=1}^t \rho_\tau} \\ &= \frac{nD^2}{2\sum_{\tau=1}^t \rho_\tau} + \frac{L^2}{K}\left(2 + C + \frac{K}{n}\right)\frac{\sum_{\tau=1}^t \rho_\tau^2}{\sum_{\tau=1}^t \rho_\tau} \\ &= \frac{DL}{2}\sqrt{\frac{n}{K}\left(2 + C + \frac{K}{n}\right)}\frac{2+\ln t}{\sqrt{t}-1} \blacksquare\end{aligned} \tag{3.45}$$

Порівняння швидкості збіжності алгоритмів (2.6), (2.13), (2.19) та їх варіантів із застосуванням згладжування для цільової функції $F(x_1, x_2) = \log(1 + x_1^2)^2 + 10x_2^2$ показано на рисунку 3.1 з кількістю ітерацій в таблиці 3.1.

Табл. 3.1: Порівняння асимптотичної швидкості збіжності методу градієнтного спуску та його модифікацій

	Градієнтний спуск	**Метод Поляка**	**Метод Нестерова**
$k = 0$	3230	292	200
$k = 7$	2472	80	55

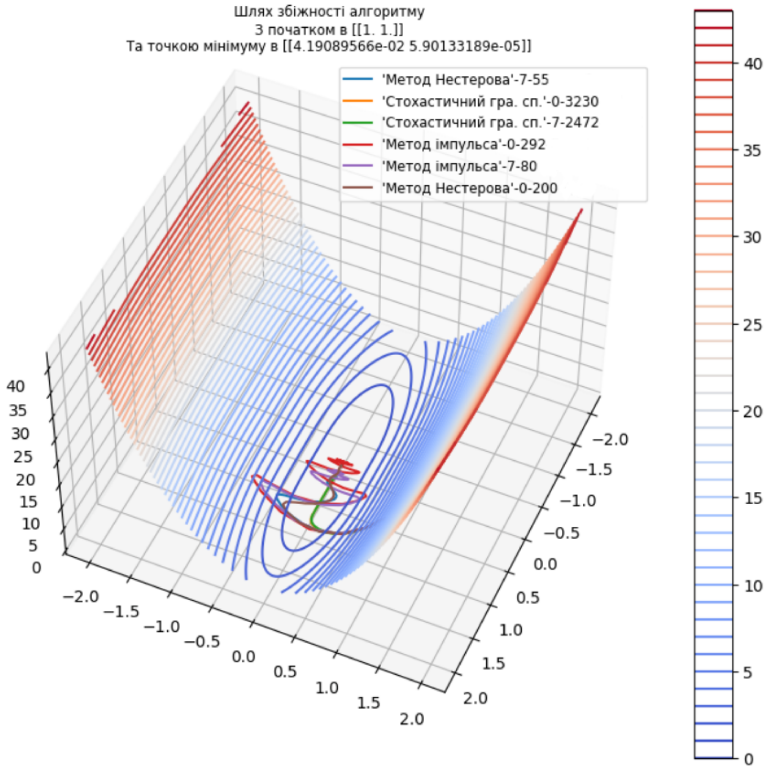

Рис. 3.1: Порівняння швидкості збіжності методів та їх аналогів з використанням згладжування за Стєкловим для цільової функції $F(x_1, x_2) = \log(1 + x_1^2)^2 + 10x_2^2$ при заданих параметрах $\rho = 0.01$, $\gamma = 0.9$ та початковій точці $(1.0, 1.0)$.

3.4 Метод Шора

Окрім стратегії нормування кроку для прискорення збіжності узагальненого градієнтного методу (3.15), як показано у рівнянні (3.20), досягнути аналогічного ефекту можливо використовуючи зміну топології цільової функції (стиснення/розтягування) під час процесу власне спуску. Даний механізм зміни топології використовується в r-алгоритмі Шора [90]. Нехай дано негладку цільову функцію $f : \mathbb{R}^n \to \mathbb{R}$. Позначимо константу $\gamma \in (0, 1)$ та матрицю перетворення $B_0 = I$, тоді рекурентна формула методу Шора має наступний вигляд:

$$\begin{array}{ll} x^{k+1} & = x^k - \rho_k B_k B_k^T \nabla f(x^k) \\ r^{k+1} & = B_k^T \left(\nabla f(x^{k+1}) - \nabla f(x^k)\right) \\ B_{k+1} & = B_k \left(I - \gamma r^{k+1} (r^{k+1})^T\right). \end{array} \tag{3.46}$$

Нехай цільова функція може бути вираженою у формі задачі оптимізації з обмеженнями у формі $x \in X = \{x \in \mathbb{R}^n : h(x) \le 0\}$, де $h(x)$ – опукла функція [31]:

$$F(x) = f(x) + M \max\{0, h(x)\} \to \min_{x \in \mathbb{R}^m} \tag{3.47}$$

де $M > 0$ – достатньо велика завчасно невідома константа. Розглянемо аналогічну до цільової функції (3.46) форму [105]:

$$F(x) = f(x) + M \left\|x - \pi_X(x)\right\|^\gamma \to \min_{x \in \mathbb{R}^m}, \qquad 0 < \gamma \le 1 \tag{3.48}$$

де $\pi_X(x) = \arg\min_{y \in X} \|y - x\|^2$ – евклідова проекція елемента на опуклу множину; для $x \in X$, очевидно, $\pi_X(x) = x$. Тут знову M має бути достатньо великим для Ліпшицевої f та $\gamma = 1$, а F не є Ліпшицевою, коли $\gamma \in (0, 1)$. Якщо множина X визначена лінійними обмеженнями, то знаходження $\pi_X(x)$ є задачею квадратичного програмування. З простими обмеженнями, наприклад, з двосторонніми обмеженнями на змінні або обмеженнями у вигляді симплекса, цю задачу можна розв'язати аналітично.

Якщо цільову функцію (3.48) обмежити на опуклій множині X, то її можливо звести до аналогічної задачі безумовної оптимізації:

$$F(x) = f\left(\pi_X(x)\right) + M \left\|x - \pi_X(x)\right\|^\gamma \to \min_{x \in \mathbb{R}^m} \tag{3.49}$$

де $\gamma > 0$ та $M > 0$ – довільні додатні числа. Функція $F(x)$ в є негладкою, а при $\gamma = 1$ вона є Ліпшицевою. Вона може бути неопуклою, але не має локальних мінімумів,

відмінних від розв'язку задачі. Перевага форми відносно та полягає в тому, що параметри штрафу M та γ в можуть бути вибрані довільно, але в M їх необхідно вибирати ретельно. Для знаходження субградієнта $g_F(x) = g_f(x) + M\left(x - \pi_X(x)\right)/\|x - \pi_X(x)\|$ при $\gamma = 1$ ми припускаємо, що $g_F(x) = 0$, $\pi_X(x) = x$.

Для демонстрації швидкості збіжності, проведемо оптимізацію цільової функції $f(x) = \sum_{i=1}^{n}(1.2)^{i-1}|x_i - 1|,\ n \in [10, 100]$, за умов обмежень $X = \{x \in \mathbb{R}^n : x_i \in [c_i, d_i], i = 1, \dots, n\}$. Розв'язок задачі: $x_i^* \in \max\{c_i, \min\{1, d_i\}\}$, $i = 1, \dots, n$. Якщо $1 \in [c_i, d_i]$, то $x^* = (1, 1, \dots, 1)^T$, $f^* = f^*(x^*) = 0$. У цьому випадку проекція $\pi_X(x)$ точки x на X знаходиться аналітично, $(\pi_X(x))_i = \max\{c_i, \min\{x_i, d_i\}\}$, $i = 1, \dots, n$. Штрафний коефіцієнт M теоретично може бути довільним; у числових експериментах він змінювався в межах $[10^1, 10^4]$, параметр γ дорівнював 1. Результати числових експериментів представлені в таблиці 3.2, де параметр n позначає розмірність простору, M – коефіцієнт штрафу. Комірки містять інформацію про точність ε або δ, кількість ітерацій r-алгоритму itn та час розв'язання в секундах.

Табл. 3.2: Тестування методу точного штрафу за умов обмежень: $x_i \in [0, 1]$, $i = 1, \dots, n$

	$M = 1$	$M = 10000$
$n = 10$	$Iter = 117, t_{sec} = 0.0732$	$Iter = 199, t_{sec} = 0.1904$
$n = 30$	$Iter = 381, t_{sec} = 0.1336$	$Iter = 3579, t_{sec} = 2.5173$
$n = 100$	$Iter = 1704, t_{sec} = 0.6013$	$Iter = 2420, t_{sec} = 1.0253$

У таблиці 3.3 наведено результати розв'язання задачі методом штрафів за обмежень: $\sum_{i=1}^{n} x_i \le n/2$, $x_i \in [0, 1]$, $i = 1, \dots, n$. У цьому випадку значення M є важливим і має бути вибране достатньо великим. Пусті клітинки означають, що r-алгоритм зупинився в недосяжній точці.

Табл. 3.3: Тестування методу точного штрафу (за обмежень: $\sum_{i=1}^{n} x_i \le n/2$, $x_i \in [0, 1]$, $i = 1, \dots, n$.)

	$M = 1$	$M = 100$	$M = 10000$
$n = 10$	–	$\delta = 6.821032\text{e-}04$ $Iter = 1000$ $t_{sec} = 4.7020$	$\delta = 6.454530\text{e-}04$ $Iter = 1000$ $t_{sec} = 4.7005$
$n = 30$	–	$\delta = 3.508656\text{e-}04$ $Iter = 487$ $t_{sec} = 0.7939$	$\delta = 2.082616\text{e-}03$ $Iter = 1000$ $t_{sec} = 2.6072$
$n = 50$	–	–	$\delta = 7.590022\text{e-}03$ $Iter = 804$ $t_{sec} = 1.2983$

У таблицях 3.4, 3.5 наведено результати розв'язання задачі методом штрафів за умов блокових обмежень: $x_i \in [0, 1]$, $i = 1, \dots, n$, та за умов обмежень $\sum_{i=1}^{n} x_i \leq n/2$, $x_i \in [0, 1]$, $i = 1, \dots, n$ відповідно. У цих прикладах значення M не відіграє суттєвої ролі в обчисленнях, але час виконання є великим порівняно з таблицями 3.2, 3.3. Це пояснюється трудомісткою оцінкою узагальнених градієнтів штрафної функції методом скінченних різниць. Якби скінченні різниці обчислювалися паралельно, то час виконання потенційно зменшився б на n і був би порівнянним з даними з таблиць 3.2, 3.3.

Табл. 3.4: Тестування методу точної проективної штрафної функції (обмеження: $x_i \in [0, 1]$, $i = 1, \dots, n$).

	$M = 1$	$M = 100$	$M = 10000$
$n = 10$	$\varepsilon = 3.296448\text{e-}08$ $Iter = 163$ $t_{sec} = 2.6436$	$\varepsilon = 1.516709\text{e-}08$ $Iter = 214$ $t_{sec} = 1.4820$	$\varepsilon = 5.581399\text{e-}09$ $Iter = 215$ $t_{sec} = 1.7683$
$n = 30$	$\varepsilon = 1.049490\text{e-}07$ $Iter = 605$ $t_{sec} = 15.6210$	$\varepsilon = 2.446992\text{e-}06$ $Iter = 441$ $t_{sec} = 7.9904$	$\varepsilon = 4.211114\text{e-}03$ $Iter = 5000$ $t_{sec} = 175.070$
$n = 80$	–	$\varepsilon = 2.555098\text{e-}05$ $Iter = 1667$ $t_{sec} = 148.264$	$\varepsilon = 8.120949\text{e-}04$ $Iter = 1427$ $t_{sec} = 125.728$

Табл. 3.5: Тестування методу точного штрафу (за обмежень: $\sum x_i \leq n/2$, $x_i \in [0, 1]$, $i = 1, \dots, n$)

	$M = 1$	$M = 10000$
$n = 10$	$\delta = 2.888739\text{e-}04$ $Iter = 157$ $t_{sec} = 4.4791$	$\delta = 2.888739\text{e-}04$ $Iter = 166$ $t_{sec} = 4.6442$
$n = 30$	$\delta = 5.652905\text{e-}04$ $Iter = 1000$ $t_{sec} = 108.9862$	$\delta = 5.652905\text{e-}04$ $Iter = 633$ $t_{sec} = 51.0818$

3.5 Висновки до розділу

Отже, у даному розділі було показано спосіб узагальнення класичного методу градієнтного спуску (2.6) для негладких цільових функцій за допомогою субградієнтної оптимізації (3.15), (3.27) та множин Кларка. Завдяки властивостям субградієнта, розглянуту модифікацію можна застосувати як до класичних градієнтних методів (2.13), (2.19), так і до адаптивних методів (2.48), (2.52), (2.58) для знаходження оптимальних значень моделей з великою кількістю параметрів, такі як моделі DL з нелінійними функціями активації [63,123]. Розглянуті властивості, як (3.12), дозволяють обчислювати складні цільові функції, градієнт яких немає аналітичної форми, наприклад функції $\min f(x)$. Даний клас функцій використовується в запропонованому методі SQC. Стохастична апроксимація може бути застосована до узагальненого градієнтного методу (2.40) для оптимізації моделі на великих даних.

Визначивши методи оптимізації, узагальнені на негладку цільову функцію, та додаткові властивості існування екстремумів для негладких функції, покажемо запропонований метод SQC у наступному розділі.

РОЗДІЛ 4. МЕТОД СТОХАСТИЧНОЇ КВАЗІГРАДІЄНТНОЇ КЛАСТЕРИЗАЦІЇ

У цьому розділі представлено метод SQC – альтернативний підхід до задачі стохастичної квазіградієнтної кластеризації дискретних ймовірнісних розподілів, що базується на мінімізації метрики Канторовича–Рубінштейна (відстані Васерштейна) між розподілом фіксованих елементів та розподілом рухомих центрів [64]. Спочатку сформульовано задачу квантизації як транспортну задачу з рухомими центрами, де одночасно оптимізуються позиції центрів, їхні ймовірності та розподіл транспортних потоків між елементами і центрами. Показано ключову особливість отриманої задачі – нелінійність цільової функції за позиціями центрів при лінійності за транспортними змінними, – що дозволяє декомпозицію на послідовність підзадач, кожна з яких є задачею лінійного програмування, зберігаючи монотонне зменшення відстані між фіксованим та апроксимуючим розподілами.

Далі здійснено зведення задачі квантизації до задачі неопуклого стохастичного програмування шляхом аналітичного усунення оптимізації за транспортними змінними та ймовірностями центрів [64]. На основі стохастичного формулювання побудовано рекурентне правило оновлення позицій центрів алгоритму SQC та його адаптивні модифікації на основі адаптивних градієнтних методів (2.48), (2.52), (2.58). Під час опису було зроблено акцент на перевагах стохастичного підходу щодо вимог до пам'яті обчислювального пристрою, оскільки метод дає можливість оновлювати кожен центроїд лише з використанням одного випадково обраного елемента $\tilde{\xi}$ замість усієї множини вхідних даних $\{\xi_1, \dots, \xi_N\}$. Іншими словами, метод пропонує кращу просторову складність $O(1)$, замість лінійної просторової складності $O(N)$, що притаманна методу K-середніх. Особливу увагу приділено задачі побудови початкового наближення центроїдів [110], для якої запропоновано комбінований підхід, що поєднує жадібний алгоритм попереднього відбору кандидатів із задачею ЗЦЛП з обмеженням кардинальності [44], а також проведено порівняльний аналіз із імовірнісною стратегією ініціалізації K-середніх++ [5] на відкритих наборах даних [58]. Отримано верхню

оцінку швидкості збіжності методу SQC, що ґрунтується на теорії узагальнено-диференційованих функцій та забезпечує збіжність до зв'язної компоненти множини стаціонарних точок з ймовірністю один.

Зрештою, продемонстровано прикладне застосування методу SQC до задачі колірного квантизації зображень [94] – одного з підходів до стиснення зображень із втратами, що зберігає оригінальну роздільну здатність при суттєвому зменшенні обсягу інформації. Задачу кольорової квантизації інтерпретовано як неопуклу стохастичну транспортну задачу [64], де пікселі зображення є фіксованими елементами у тривимірному колірному просторі RGB, а кольори оптимальної палітри визначаються як рухомі центри. Проведено серію чисельних експериментів зі стиснення зображень різної роздільної здатності на підмножині набору даних ImageNet [108], що підтверджує масштабованість та ефективність запропонованого підходу порівняно з існуючими методами кластеризації (1.8), (1.13), (1.20) і (1.26). Запропонований метод стохастичної квантизації не лише усуває обмеження існуючих методів квантизації, пов'язані з масштабованістю та лінійною залежністю складності від кількості елементів вхідної множини, але й забезпечує теоретичні гарантії збіжності, що мотивує його застосування як для класичних задач розміщення об'єктів обслуговування та апроксимації розподілів, так і для сучасних задач стиснення даних у середовищах із обмеженими обчислювальними ресурсами.

4.1 Транспортна задача з рухомими центрами

Отже, задача стохастичної квазіградієнтної кластеризації може бути зведена до задачі наближення одного дискретного імовірнісного розподілу з великою кількістю атомів іншим дискретним розподілом із суттєво меншою кількістю атомів. Якщо попит споживачів є фіксованим, то задача розміщення об'єктів обслуговування зводиться до апроксимації дискретного розподілу клієнтів дискретним розподілом із меншим числом центрів обслуговування [64]. Запропонований підхід розглядає задачу квантизації як мінімізацію відстані Васерштейна (метрики Канторовича–Рубінштейна) між розподілом фіксованих елементів та розподілом рухомих центрів, що дозволяє одночасно оптимізувати як позиції центрів, так і розподіл транспортних потоків між елементами та центрами.

Формально визначимо метрику Канторовича–Рубінштейна для двох скінченних багатовимірних дискретних розподілів. Нехай Ξ та Z є дискретними випадковими векторами з атомами $\{\xi_1, \dots, \xi_N\}$, $\{z_1, \dots, z_K\}$ та відповідними ймовірностями $\{p_1, \dots, p_N\}$, $\{q_1, \dots, q_K\}$. Нехай c_{ik} є невід'ємною відстанню між точками ξ_i та z_k. Метрика Канторовича–Рубінштейна між розподілами визначається як оптимальне значення наступної транспортної задачі [64]:

$$D(\Xi, Z) = \min_{x_{ik}} \sum_{i=1}^{N} \sum_{k=1}^{K} c_{ik}\, x_{ik} \tag{4.1}$$

за обмежень:

$$\begin{array}{ll} \sum_{k=1}^{K} x_{ik} = p_i, & \sum_{i=1}^{N} x_{ik} = q_k, \\ x_{ik} \geq 0, & i = 1, \dots, N, \quad k = 1, \dots, K. \end{array} \tag{4.2}$$

У термінології задач розміщення, формулювання (4.1)-(4.2) відповідає транспортній складовій задачі розміщення-розподілу, де величини p_i та q_k є нормованими обсягами постачання та споживання відповідно, а змінні x_{ik} є нормованими обсягами транспортування [64]. Обмеження (4.2) описують баланс між постачанням та споживанням, з якого випливає рівність $\sum_{i=1}^{N} p_i = \sum_{k=1}^{K} q_k$. Відповідно, система лінійних рівнянь (4.2) є сумісною, але не незалежною, що надає певну гнучкість у формулюванні задачі, зокрема дозволяє замінити одну з систем рівностей нерівностями [64].

Розглянемо тепер центральну задачу квантизації як задачу наближення розподілу з рухомими центрами. Нехай задано d-вимірний дискретний розподіл з фіксованими атомами $\Xi = \{\xi_i,\, i = 1, \dots, N\}$, де $\xi_i \in \mathbb{R}^d$, та відповідними ймовірностями $p_i > 0$, . Метою є наближення цього розподілу іншим дискретним розподілом із рухомими центрами $Y = \{y_k,\, k = 1, \dots, K\}$, де $y_k \in Y \subseteq \mathbb{R}^d$ та $K \ll N$, з відповідними ймовірностями $q_k \geq 0$, $\sum_{k=1}^{K} q_k = 1$. Координати центрів y_k, ймовірності q_k та транспортні змінні x_{ik} є змінними задачі оптимізації. Позиції рухомих центрів із найменшою відстанню Канторовича–Рубінштейна до заданого розподілу Ξ визначаються розв'язком наступної задачі оптимізації:

$$\min_{\substack{y=\{y_1,\dots,y_K\}\in Y^K \subset \mathbb{R}^{dK} \\ q=\{q_1,\dots,q_K\}\in \mathbb{R}_+^K \\ x=\{x_{ik}\geq 0\}}} \sum_{i=1}^{N} \sum_{k=1}^{K} \operatorname{dist}(\xi_i, y_k)^r\, x_{ik} \tag{4.3}$$

за обмежень:

$$\sum_{k=1}^{K} x_{ik} = p_i, \ \sum_{k=1}^{K} q_k = 1, \quad i = 1, \dots, N, \tag{4.4}$$

де $\mathrm{dist}(\xi_i, y_k)$ є метрикою на просторі $\mathbb{R}^d$, а $r \geq 1$ є параметром, що визначає порядок відстані Васерштейна. Зв'язок між транспортними змінними x_{ik} та ймовірностями центрів q_k встановлюється співвідношенням $q_k = \sum_{i=1}^{N} x_{ik}$, що визначає ймовірність кожного центру як сумарний обсяг транспортування до нього. Друге обмеження нормування з (4.4) є наслідком першого обмеження та умови $\sum_{i=1}^{N} p_i = 1$, оскільки:

$$\sum_{k=1}^{K} q_k = \sum_{k=1}^{K} \sum_{i=1}^{N} x_{ik} = \sum_{i=1}^{N} \sum_{k=1}^{K} x_{ik} = \sum_{i=1}^{N} p_i = 1 \tag{4.5}$$

проте наведено його явно для повноти опису ймовірнісної структури апроксимуючого розподілу.

Задача (4.3)-(4.4) полягає в наступному: кожен фіксований елемент ξ_i з ймовірністю p_i повинен бути повністю «розподілений» між центрами y_k через транспортні змінні x_{ik}, а відстань $\mathrm{dist}(\xi_i, y_k)^r$ визначає вартість переміщення одиниці ймовірнісної маси від елемента ξ_i до центру y_k. Цільова функція є зваженою сумою відстаней між елементами та центрами, де кожна відстань зважена відповідним обсягом транспортування. Зазвичай для визначення відстані між точками у цільовій функції (4.3) використовується l_p-норма:

$$\mathrm{dist}(\xi_i, y_k) = \left(\sum_{l=1}^{d} |\xi_{il} - y_{kl}|^p \right)^{1/p}, \quad p \geq 1 \tag{4.6}$$

де ξ_{il} та y_{kl} позначають l-ту координату елемента ξ_i та центру y_k відповідно. Випадок $p = 2$ відповідає евклідовій нормі, яка є найбільш поширеною у прикладних задачах розміщення. Задача (4.3) може бути класифікована як задача K-медіани із зваженою відстанню та неперервними позиціями для K центрів, що встановлює зв'язок між задачами кластеризації, планування розміщення та апроксимації розподілів. Вибір функції відстані є визначальним як для формулювання задачі, так і для методів її розв'язання.

Ключовою особливістю задачі (4.3)-(4.4) є її нелінійність та некомпактність: цільова функція є нелінійною за позиціями центрів y_k, а мінімізація здійснюється одночасно за неперервними змінними y_k та транспортними змінними x_{ik}. Це унеможливлює пряме застосування стандартних методів лінійного програмування. Проте задача допускає ефективну декомпозицію на послідовність підзадач, кожна з

яких є або задачею лінійного програмування, або задачею безумовної нелінійної оптимізації. Алгоритм розв'язання складається з двох етапів:

Етап 1 (побудова початкового наближення). Розглядається задача (4.3)-(4.4) з припущенням, що центри y_k розташовані лише в точках фіксованих елементів ξ_i. Ця задача формулюється як задача цілочислового програмування з обмеженням на кардинальність (кількість активних центрів не повинна перевищувати K):

$$\min_{\delta_i, x_{ij}} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathrm{dist}(\xi_i, \xi_j)^r \, x_{ij} \tag{4.7}$$

за обмежень:

$$\begin{array}{llll} \sum_{i=1}^{N} x_{ij} = p_j, & \sum_{j=1}^{N} x_{ij} \le \delta_i, & & i = 1, \dots, N \\ \sum_{i=1}^{N} \delta_i \le K, & \delta_i \in \{0,1\}, x_{ij} \ge 0, & & j = 1, \dots, N \end{array} \tag{4.8}$$

де δ_i є технічними булевими змінними, що формалізують обмеження кардинальності на кількість активних центрів. Розв'язок задачі (4.7)-(4.8) визначає K елементів ξ_i із $\delta_i = 1$, які слугують початковим наближенням для позицій рухомих центрів у задачі (4.3).

Етап 2 (ітераційне уточнення). Задача (4.7)-(4.8) розв'язується наближено шляхом розв'язання послідовності пар оптимізаційних підзадач. Розв'язання кожної пари підзадач монотонно зменшує відстань між фіксованим та апроксимуючим розподілами. Ітерації припиняються, коли відстань перестає зменшуватись. Підзадачі формулюються наступним чином.

Підзадача 1 (оптимізація позицій центрів). За фіксованих значень транспортних змінних x_{ik}, отриманих на попередній ітерації, розв'язується задача:

$$\min_{y_k} \sum_{i=1}^{N} \sum_{k=1}^{K} \mathrm{dist}(\xi_i, y_k)^r \, x_{ik} \tag{4.9}$$

що є задачею, в якій оновлюються позиції K центрів y_k при збереженні фіксованого розподілу транспортних потоків. Ця задача розпадається на K незалежних підзадач оптимізації позиції кожного центру.

Підзадача 2 (оптимізація транспортних змінних). За фіксованих позицій центрів y_k, отриманих із підзадачі 1, розв'язується задача лінійного програмування:

$$\min_{x_{ik}} \sum_{i=1}^{N} \sum_{k=1}^{K} \mathrm{dist}(\xi_i, y_k)^r \, x_{ik} \tag{4.10}$$

за обмежень (4.4) та $x_{ik} \geq 0$. Розв'язок цієї задачі має характерну структуру: для кожного елемента $j = 1, \dots, N$ існує індекс k_j такий, що $x_{k_j j} = p_j$ та $x_{kj} = 0$ для $k \neq k_j$, тобто кожен елемент повністю приписується до найближчого центру.

Монотонне спадання цільової функції задачі (4.3) при почерговому розв'язанні підзадач обґрунтовується трьома фактами:

1. обидві підзадачі мають спільну цільову функцію;
2. відстань зменшується при розв'язанні кожної підзадачі;
3. розв'язок однієї підзадачі використовується як початкова точка для наступної.

Скінченність процесу оптимізації випливає з того, що оптимальний вектор x_{ik} підзадачі 2 має описану вище структуру повного приписування, кількість таких комбінацій є скінченною, а процес монотонно зменшує цільову функцію. Ефективність та швидкість збіжності запропонованого алгоритму можна порівняти з ефективністю алгоритму K-середніх, оскільки обидва алгоритми використовують швидку побудову початкового наближення та двокрокову процедуру його уточнення, а швидкість збіжності до локального оптимуму оцінюється як $O(\log K)$. Підсумовуючи, задача (4.7)-(4.8) поєднує три взаємопов'язані аспекти стохастичної квазіградієнтної кластеризації:

1. оптимізація позицій рухомих центрів $y_k \in Y \subseteq \mathbb{R}^d$, що визначає геометричне розташування апроксимуючого розподілу;
2. оптимізація ймовірностей q_k центрів, що задає вагову структуру апроксимуючого розподілу та визначається як $q_k = \sum_{i=1}^{N} x_{ik}$;
3. оптимізація транспортних змінних x_{ik}, що визначає оптимальний план перерозподілу ймовірнісної маси від фіксованих елементів до рухомих центрів.

Декомпозиційний алгоритм етапу 2 зводить цю складну задачу спільної оптимізації до послідовності добре досліджених підзадач, зберігаючи обчислювальну ефективність та гарантуючи збіжність до локального оптимуму за скінченне число кроків.

4.2 Зведення задачі до стохастичної форми

Отже, декомпозиційний алгоритм етапу 2, описаний у попередньому підрозділі, забезпечує збіжність до локального оптимуму за скінченне число кроків шляхом почергового розв'язання підзадачі оптимізації позицій центрів та підзадачі лінійного програмування для транспортних змінних. Проте для великих значень N кожна ітерація підзадачі 2 потребує одночасного зберігання та обробки всіх N фіксованих елементів $\{\xi_i\}_{i=1}^N$ для обчислення відстаней $\operatorname{dist}(\xi_i, y_k)^r$, що призводить до просторової складності $O(Nd)$ за обсягом оперативної пам'яті. У прикладних задачах, де кількість фіксованих елементів N може сягати порядку 10^5-10^7, ця вимога стає обмеженням масштабованості, оскільки повна вибірка $\{\xi_i\}$ не завжди може бути завантажена в оперативну пам'ять обчислювального пристрою [64]. Альтернативний підхід полягає у зведенні задачі (4.3)-(4.4) до задачі неопуклого стохастичного програмування, що допускає ефективне наближене розв'язання методами стохастичної апроксимації з використанням лише одного елемента в пам'яті на кожній ітерації, тобто з просторовою складністю $O(d)$ незалежно від N.

Ключовим спостереженням, що обґрунтовує подібне зведення, є структура оптимального розв'язку внутрішньої задачі мінімізації за транспортними змінними x_{ik} при фіксованих позиціях центрів y_k. Розглянемо задачу (4.3) за обмеження (4.4) для фіксованого елемента ξ_i: необхідно мінімізувати $\sum_{k=1}^K \operatorname{dist}(\xi_i, y_k)^r x_{ik}$ за умови $\sum_{k=1}^K x_{ik} = p_i$, $x_{ik} \geq 0$. Оскільки ця задача є лінійною за x_{ik} з обмеженням на суму, її оптимальний розв'язок зосереджений в одній вершині симплексу, тобто вся ймовірнісна маса p_i приписується до найближчого центру. Формально, для кожного i існує індекс $k^*(i) = \arg\min_{1 \leq k \leq K} \operatorname{dist}(\xi_i, y_k)^r$ такий, що $x_{ik^*(i)} = p_i$ та $x_{ik} = 0$ для $k \neq k^*(i)$. Підставляючи цей оптимальний розв'язок у цільову функцію (4.3), отримаємо послідовне зведення задачі:

$$\min_{\substack{y \in Y^K \\ q \in \mathbb{R}_+^K \\ x_{ik} \geq 0}} \sum_{i=1}^N \sum_{k=1}^K \operatorname{dist}(\xi_i, y_k)^r \, x_{ik} \overset{\sum_k x_{ik} = p_i}{=} \min_{y \in Y^K} \sum_{i=1}^N p_i \min_{1 \leq k \leq K} \operatorname{dist}(\xi_i, y_k)^r \tag{4.11}$$

Таким чином, оптимізація за транспортними змінними x_{ik} та ймовірностями q_k усувається аналітично, і задача (4.3)-(4.4) зводиться до задачі безумовної негладкої та неопуклої оптимізації:

$$\min_{y=(y_1,\dots,y_K)\in Y^K\subset\mathbb{R}^{dK}} F(y_1,\dots,y_K) \tag{4.12}$$

де цільова функція F визначається як:

$$F(y)=\sum_{i=1}^{N} p_i \min_{1\le k\le K} \operatorname{dist}(\xi_i,y_k)^r = \mathbb{E}_{i\sim p}\left[\min_{1\le k\le K}\operatorname{dist}(\xi_i,y_k)^r\right] \tag{4.13}$$

Представлення цільової функції (4.13) у формі математичного сподівання відносно дискретного розподілу з ймовірностями $\{p_i\}$ визначає задачу (4.13) як задачу стохастичного програмування, в якій мінімізується очікувана вартість приписування випадково обраного елемента ξ_i до найближчого центру [64]. Ця інтерпретація відкриває можливість застосування методів стохастичної апроксимації, що використовують на кожній ітерації лише один випадково обраний елемент замість усієї сукупності $\{\xi_1,\dots,\xi_N\}$.

Функція $F(y)$ є неопуклою внаслідок наявності оператора мінімуму за індексом центру k, що унеможливлює гарантію збіжності до глобального оптимуму. Водночас функція $F(y)$ є субдиференційовною, що є достатнім для застосування методів субградієнтного спуску. Обчислимо субградієнт $\partial F(y)$ за позиціями центрів. Використовуючи евклідову норму як метрику, тобто $\operatorname{dist}(\xi_i,y_k)=\|y_k-\xi_i\|$, та теорему 15 про субдиференціювання мінімуму, отримаємо:

$$\partial F(y)=\sum_{i=1}^{N} p_i\,\partial \min_{1\le k\le K}\|y_k-\xi_i\|^r = \sum_{i=1}^{N} p_i\,\partial\|y_{k^*}-\xi_i\|^r,\quad k^*\in S(\xi_i,y) \tag{4.14}$$

де $S(\xi_i,y)=\arg\min_{1\le k\le K}\|y_k-\xi_i\|^r$ є множиною індексів найближчих центрів до елемента ξ_i. Визначимо розбиття Вороного множини елементів $\{\xi_1,\dots,\xi_N\}$ відносно поточних позицій центрів: позначимо $I_k=\{i: k\in S(\xi_i,y)\}$ множину індексів елементів, для яких центр y_k є найближчим. Тоді субградієнт $\partial F/\partial y_k$ за позицією K-го центру обчислюється як:

$$\frac{\partial F}{\partial y_k}=\sum_{i\in I_k} p_i\cdot r\,\|y_k-\xi_i\|^{r-2}(y_k-\xi_i),\quad k=1,\dots,K \tag{4.15}$$

Інтуїтивна інтерпретація виразу (4.15) полягає в наступному: субградієнт за положенням центру y_k є зваженою сумою напрямків від y_k до його належних елементів ξ_i, де кожен напрямок масштабується ймовірністю p_i та ступеню відстані $\|y_k-\xi_i\|^{r-2}$. Параметр r відіграє роль регулятора згладжування: при $r=2$ градієнт $2(y_k-\xi_i)$ є лінійним за відхиленням і не залежить від абсолютної відстані; при

$r > 2$ далекі елементи вносять пропорційно більший вклад у субградієнт, що забезпечує більш виражене «притягування» центрів до віддалених елементів.

На основі стохастичної оцінки субградієнта (4.15) побудуємо рекурентне правило оновлення позицій центрів. Класичний підхід стохастичної апроксимації полягає в заміні повного субградієнта його незміщеною стохастичною оцінкою, обчисленою за одним випадково вибраним елементом $\tilde{\xi} \sim \{p_i\}$. Нехай на ітерації t обрано елемент $\tilde{\xi}$ та визначено індекс найближчого центру $k^* \in S(\tilde{\xi}, y^t)$. Стохастична оцінка субградієнта для центру y_k визначається як:

$$g_k^t = \begin{cases} r\,\|\tilde{\xi} - y_k^t\|^{r-2}(y_k^t - \tilde{\xi}), & k = k^*, \\ 0, & k \neq k^*. \end{cases} \tag{4.16}$$

Отже, на кожній ітерації оновлюється лише позиція одного центру – найближчого до обраного елемента, – тоді як решта центрів залишаються незмінними. Використовуючи оцінку (4.16), отримаємо рекурентне правило оновлення у загальному вигляді. Для пакетного варіанту, що використовує всі елементи множини I_k^t (множини елементів, приписаних до центру k на ітерації t), оновлення має вигляд:

$$y_k^{t+1} = y_k^t - \rho_t\, g_k(y^t) = \begin{cases} y_k^t - \rho_t \dfrac{r}{I} \displaystyle\sum_{i \in I_k^t} \|y_k^t - \xi_i\|^{r-2}(y_k^t - \xi_i), & I_k^t \neq \emptyset, \\ y_k^t, & I_k^t = \emptyset, \end{cases} \tag{4.17}$$

де $\rho_t > 0$ є кроком спуску на ітерації t. Для забезпечення збіжності стохастичної апроксимації послідовність кроків повинна задовольняти класичні умови Роббінса-Монро: $\sum_{t=0}^{\infty} \rho_t = \infty$ та $\sum_{t=0}^{\infty} \rho_t^2 < \infty$, що гарантує, з одного боку, достатню сумарну довжину кроків для досягнення будь-якого локального мінімуму, а з іншого – зменшення дисперсії стохастичної оцінки. Типовим вибором є $\rho_t = \rho_0/(t+1)^{\alpha}$, при $1/2 < \alpha \leq 1$.

Повний алгоритм SQC формулюється наступним чином. Задано множину елементів $\{\xi_i,\, i = 1, \dots, N\}$ та гіперпараметри: крок спуску $\rho > 0$, кількість центроїдів K, кількість ітерацій T та параметр згладжування $r = 3$.

1. Ініціалізація центроїдів: $y^0 = \{y_k^0,\, k = 1, \dots, K\}$.
2. Для кожної ітерації $t = 0, 1, \dots, T-1$:
   i. вибір випадкового елемента $\tilde{\xi} \in \{\xi_i,\, i = 1, \dots, N\}$ з ймовірностями $\{p_i\}$;
   ii. пошук найближчого центру: $k^* \in S(\tilde{\xi}, y^t) = \arg\min_{1 \leq k \leq K} \|\tilde{\xi} - y_k^t\|$;

iii. обчислення стохастичної оцінки субградієнта: $g^t_{k^*} = r\,\|\tilde{\xi} - y^t_{k^*}\|^{r-2}(y^t_{k^*} - \tilde{\xi})$;

iv. оновлення позиції обраного центру: $y^{t+1}_{k^*} = \Pi_Y(y^t_{k^*} - \rho_t\, g^t_{k^*})$.

3. Позначити $y^T = \{y^T_k,\, k = 1, \dots, K\}$.

Перевагою стохастичного підходу порівняно з декомпозиційним алгоритмом етапу 2 є зменшення вимог до оперативної пам'яті обчислювального пристрою: на кожній ітерації алгоритм SQC оперує лише одним випадково вибраним елементом $\tilde{\xi}$ та K позиціями центрів, що відповідає просторовій складності $O((K+1)d) \sim O(d)$, тоді як детерміністичний алгоритм потребує одночасної присутності в пам'яті всієї сукупності $\{\xi_1, \dots, \xi_N\}$ з просторовою складністю $O(Nd)$. Ця властивість є суттєвою перевагою при великих обсягах вхідних даних, коли повна вибірка не вміщується в оперативну пам'ять.

Зазначимо, що алгоритм SQC потребує знаходження найближчого кластерного центру $\arg\min \|y_k - \xi_i\|$ для кожної точки $\{\xi_i,\, i = 1, \dots, N\}$. Ця операція може бути обчислювально затратною при великих N. Більше того, якщо точки ξ_i послідовно генеруються з неперервного розподілу, вибірка $\{\xi_i,\, i = 1, 2, \dots\}$ може бути потенційно необмеженою. Стохастичний алгоритм (4.16) вирішує цей недолік, використовуючи лише один обраний елемент $\tilde{\xi}^t$ на кожній ітерації для обчислення $\arg\min \|y_k - \tilde{\xi}^t\|$ на ітерації t. Проте, слідуючи підходу [111], замість повної множини $\Xi = \{\xi_i,\, i = 1, \dots, N\}$ можна використовувати міні-пакет із m елементів $\{\tilde{\xi}^t_{i^t_1}, \dots, \tilde{\xi}^t_{i^t_m}\}$, де $1 \le m < N$, що дозволяє збалансувати обчислювальну ефективність та якість алгоритмічних оновлень.

Альтернативна конструкція алгоритму SQC полягає у поділі ітераційного процесу на епохи довжини N із застосуванням механізму перемішування елементів на початку кожної епохи для вибору випадкових елементів $\tilde{\xi}^t$ [15,77]. За такого підходу кожна епоха гарантує, що всі елементи множини $\{\xi_1, \dots, \xi_N\}$ будуть використані рівно один раз (у випадковому порядку), що зменшує дисперсію стохастичних оцінок субградієнта порівняно з вибіркою із заміщенням та забезпечує більш рівномірне покриття простору елементів протягом кожної епохи.

Вибір параметра r у наведеному алгоритмі є визначальним для властивостей субградієнтних оцінок та загальної поведінки методу оптимізації. При $r = 1$ субградієнт $g^t_{k^*} = (y^t_{k^*} - \tilde{\xi})/\|\tilde{\xi} - y^t_{k^*}\|$ має одиничну норму незалежно від відстані, що

уповільнює збіжність для далеких елементів. При $r = 2$ субградієнт $g^t_{k^*} = 2(y^t_{k^*} - \tilde{\xi})$ є лінійним, що відповідає класичній задачі K-середніх. При $r = 3$ субградієнт $g^t_{k^*} = 3\|\tilde{\xi} - y^t_{k^*}\| \cdot (y^t_{k^*} - \tilde{\xi})$ масштабується пропорційно відстані, що забезпечує більш виражене переміщення центрів у напрямку віддалених елементів на початкових ітераціях та поступове згладжування оновлень при наближенні до оптимуму.

Особливий інтерес з точки зору стійкості представляє модель стійкої кластеризації, що відповідає розв'язанню задачі (4.7)-(4.8) з параметром $r \in [1, 2)$. Такий вибір підвищує стійкість моделі до викидів як у задачах квантизації, так і кластеризації, оскільки при $r < 2$ вплив далеких елементів на субградієнт зростає повільніше, ніж лінійно, що пом'якшує ефект аномальних спостережень. У цьому випадку стохастичні узагальнені градієнти цільової функції обчислюються за формулою (4.16), а стохастичний алгоритм кластеризації набуває вигляду правила оновлення (4.17) з відповідним значенням r, що забезпечує автоматичне зменшення вагового внеску елементів з аномально великими відстанями до відповідних належних центрів.

Проте алгоритм SQC із фіксованим або монотонно спадним кроком ρ_t зберігає спільне обмеження класичних методів стохастичного субградієнтного спуску – застосування єдиного скалярного кроку до всіх компонент вектора оновлення. У задачах квантизації з великою кількістю центрів K та високою розмірністю d простору елементів різні компоненти субградієнта можуть мати суттєво різні масштаби. Зокрема, центри, яким приписано мало елементів, отримують рідкісні, але інформативні оновлення, тоді як центри з великими розбиттями Вороного оновлюються часто із значною дисперсією субградієнтних оцінок. Ця асиметрія мотивує розробку адаптивних модифікацій алгоритму SQC, які динамічно підлаштовують крок оновлення для кожного центру окремо на основі накопиченої градієнтної інформації.

Повний алгоритм SQC-Adam формулюється наступним чином. Задано множину елементів $\{\xi_i,\, i = 1, \dots, N\}$ та гіперпараметри: крок спуску $\rho > 0$, кількість центрів K, кількість ітерацій T, параметр згладжування $r = 3$, коефіцієнти моментів $\beta_1 \in (0, 1)$ та $\beta_2 \in (0, 1)$.

1. Ініціалізація: центроїди $y^0 = \{y^0_k,\, k = 1, \dots, K\}$; вектори моментів $m^0_k = 0$, $v^0_k = 0$ для всіх k.

2. Для кожної ітерації $t = 0, 1, \dots, T-1$:
   i. вибір випадкового елемента $\tilde{\xi} \in \{\xi_i,\ i = 1, \dots, N\}$ з ймовірностями $\{p_i\}$;
   ii. пошук найближчого центру: $k^* \in S(\tilde{\xi}, y^t) = \arg\min_{1 \le k \le K} \|\tilde{\xi} - y_k^t\|$;
   iii. обчислення стохастичної оцінки субградієнта: $g_{k^*}^t = r\,\|\tilde{\xi} - y_{k^*}^t\|^{r-2}(y_{k^*}^t - \tilde{\xi})$;
   iv. оновлення оцінки першого моменту: $m_{k^*}^t = \beta_1\, m_{k^*}^{t-1} + (1-\beta_1)\, g_{k^*}^t$;
   v. оновлення оцінки другого моменту: $v_{k^*}^t = \beta_2\, v_{k^*}^{t-1} + (1-\beta_2)\,\|g_{k^*}^t\|^2$;
   vi. оновлення позиції обраного центру: $y_{k^*}^{t+1} = \Pi_Y\left(y_{k^*}^t - (\rho_t m_{k^*}^t)/(\sqrt{v_{k^*}^t + \varepsilon})\right)$.
3. Позначити $y^T = \{y_k^T,\ k = 1, \dots, K\}$.

Обчислювальна складність кожної ітерації алгоритму SQC-Adam залишається $O(Kd)$, тобто такою ж, як і для базового алгоритму SQC, з додатковою потребою $O(dK)$ пам'яті для зберігання векторів моментів m_k та скалярних величин v_k для кожного з K центрів.

Порівняно з базовим алгоритмом SQC, модифікація SQC-Adam має дві ключові переваги в контексті стохастичної квантизації. По-перше, оцінка першого моменту $m_{k^*}^t$ виконує роль згладжуючого фільтра послідовних субградієнтів: оскільки на кожній ітерації обирається лише один випадковий елемент $\tilde{\xi}$, субградієнтні оцінки мають високу дисперсію, і їхнє експоненційне усереднення дозволяє виділити стійкий напрямок переміщення центру, подібно до імпульсного прискорення у класичних методах оптимізації. По-друге, оцінка другого моменту $v_{k^*}^t$ адаптує крок оновлення для кожного центру окремо: центри з великою варіабельністю субградієнтів автоматично отримують менший ефективний крок, що стабілізує процес оптимізації, тоді як центри з малою варіабельністю зберігають достатню швидкість оновлення.

Додатковим підходом до стабілізації збіжності є застосування усереднення по траєкторії [15,77] до послідовностей, що генеруються пакетним правилом оновлення (4.17) та його стохастичним аналогом (4.16):

$$\begin{aligned} \bar{y}_k^{t+1} &= (1-\sigma_{t+1})\bar{y}_k^t + \sigma_{t+1} y_k^{t+1}, \\ \sigma_{t+1} &= \frac{\rho_{t+1}}{\sum_{s=1}^{t+1}\rho_s}, \quad k = 1, \dots, K, \quad t = 0, 1, \dots . \end{aligned} \tag{4.18}$$

Усереднена послідовність $\{\bar{y}_k^t\}$ є зваженою середньою попередніх позицій центрів із вагами, пропорційними крокам спуску ρ_t, що ефективно згладжує стохастичні флуктуації траєкторії. Умови збіжності для такої усередненої послідовності,

досліджені у [76], допускають вибір кроку навчання ρ_t пропорційного $1/\sqrt{t+1}$. Подібний підхід для послідовностей, що генеруються алгоритмом SQC, полягає в усередненні по кількості елементів N_k, приписаних до відповідного центру:

$$\tilde{y}_k^{t+1} = (1 - \tilde{\sigma}_{k,\,t+1})\tilde{y}_k^t + \tilde{\sigma}_{k,\,t+1} y_k^{t+1} \tag{4.19}$$

$$\tilde{\sigma}_{k,\,t+1} = \frac{1}{N_k^{t+1}} \Bigg/ \sum_{s=1}^{t+1} \frac{1}{N_k^s}, \quad k = 1, \dots, K.$$

Варіант (4.19) враховує нерівномірність частоти оновлення різних центрів: центри з малою кількістю належних елементів N_k^t отримують більші відносні ваги $\tilde{\sigma}_{k,t+1}$ у формулі усереднення, що компенсує рідкість їхніх оновлень та забезпечує більш збалансовану збіжність усіх центрів одночасно.

Підсумовуючи, стохастична апроксимація задачі квантизації з рухомими центрами реалізує альтернативний шлях розв'язання задачі (4.3)-(4.4), що базується на трьох послідовних спрощеннях:

1. аналітичне усунення оптимізації за транспортними змінними x_{ik} та ймовірностями q_k шляхом зведення задачі до мінімізації неопуклої цільової функції $F(y) = \mathbb{E}_{i\sim p}[\min_k \operatorname{dist}(\xi_i, y_k)^r]$ лише за позиціями центрів;
2. заміна повного субградієнта $\partial F(y)$, що потребує одночасного зберігання в пам'яті всіх N елементів, його стохастичною оцінкою за одним випадково вибраним елементом (або міні-пакетом із m елементів [111]), що зменшує просторову складність ітерації з $O(Nd)$ до $O(d)$, із можливістю організації ітераційного процесу в епохи з перемішуванням [14,77];
3. застосування адаптивних методів оцінювання моментів (SQC-Adam) та усереднення по траєкторії (4.18)-(4.19) для автоматичної покоординатної та поцентрової адаптації кроку оновлення, що стабілізує збіжність в умовах високої дисперсії стохастичних субградієнтних оцінок.

Ці спрощення зберігають асимптотичну збіжність до локального оптимуму задачі (4.12)-(4.13) при належному виборі послідовності кроків [76], забезпечуючи водночас практичну ефективність для великомасштабних задач квантизації та розміщення [64].

4.3 Початкова апроксимація центрів жадібним методом

Отже, як декомпозиційний алгоритм етапу 2, так і стохастичні алгоритми SQC та SQC-Adam потребують початкового наближення позицій центроїдів $y^0 = \{y_k^0,\, k = 1, \dots, K\}$. Якість цього наближення є визначальною для швидкості збіжності та якості кінцевого розв'язку, оскільки цільова функція (4.12) є неопуклою і допускає множину локальних мінімумів [110]. Зокрема, невдалий вибір початкових центроїдів може призвести до збіжності до суттєво субоптимального локального мінімуму, що знецінює переваги ітераційного уточнення на етапі 2. Класичні підходи до ініціалізації, такі як рівномірний випадковий вибір елементів із вхідної множини [35,70,71], є обчислювально ефективними, проте не забезпечують жодних гарантій щодо якості початкового наближення. Більш того, численні експериментальні дослідження свідчать про високу чутливість результатів кластеризації до випадкового вибору початкових центроїдів [18]. У цьому розділі ми послідовно розглядаємо три підходи до побудови початкового наближення: імовірнісний алгоритм K-середніх++ з теоретичними гарантіями якості, детерміністичний підхід на основі ЗЦЛП з обмеженням кардинальності, та жадібний алгоритм зменшення розмірності задачі.

Теоретично обґрунтованим покращенням стратегії рівномірного випадкового вибору є алгоритм K-середніх++ [5,84]. Цей підхід замінює рівномірне випадкове обрання центроїдів імовірнісною стратегією ініціалізації, спрямованою на ефективний розподіл початкових центроїдів у просторі елементів. Перший центроїд y_1^0 обирається рівномірно випадково з множини фіксованих елементів $\{\xi_i,\, i = 1, \dots, N\}$. Далі, для $s = 2, \dots, K$, кожний наступний центроїд y_s^0 обирається з решти елементів з ймовірністю, пропорційно квадрату відстані від елемента до найближчого з уже обраних центроїдів. Формально, ймовірність вибору елемента ξ_j як s-го центроїда визначається як [5]:

$$p_j = \frac{\min_{1 \le s \le k} \|\xi_j - y_s^0\|^2}{\sum_{i=1}^{N} \min_{1 \le s \le k} \|\xi_i - y_s^0\|^2} \tag{4.20}$$

де мінімум береться за всіма вже обраними центроїдами. Інтуїтивна інтерпретація правила (4.20) полягає в наступному: елементи, розташовані далеко від множини вже обраних центроїдів, мають суттєво вищу ймовірність бути обраними як нові

центри, що сприяє рівномірному покриттю простору елементів. Після ініціалізації центроїдів $\{y_1^0, \dots, y_K^0\}$ за цією процедурою виконується стандартна ітераційна процедура уточнення до збіжності.

Стратегія ініціалізації K-середніх++ забезпечує строгі теоретичні гарантії якості початкового наближення. Зокрема, очікуване значення цільової функції $\mathbb{E}[\sum_i \min_k \|\xi_i - y_k^0\|^2]$ після ініціалізації за процедурою (4.20) не перевищує $8(\ln K + 2)$ значення глобального оптимуму, що забезпечує $O(\log K)$ асимптотичну апроксимацію [5]. Цей теоретичний результат є суттєвим покращенням порівняно з необмеженою похибкою найгіршого випадку рівномірної випадкової ініціалізації. Емпіричні дослідження підтверджують, що стратегія K-середніх++ типово забезпечує більш компактні кластери та прискорює збіжність ітераційного процесу, оскільки початкові центроїди розташовуються у сприятливих позиціях простору елементів [5].

Проте, незважаючи на зазначені теоретичні гарантії, алгоритм K-середніх++ не забезпечує оптимальної ініціалізації центроїдів в кожному конкретному запуску, що є важливим для гарантії збіжності до глобального оптимуму неопуклої цільової функції (4.12) [110]. Ця обставина мотивує розробку детерміністичного підходу, що формулює задачу ініціалізації як задачу ЗЦЛП, в якій початкові центроїди визначаються шляхом мінімізації відстані Канторовича–Рубінштейна з обмеженням на кількість активних центрів.

Задача побудови початкового наближення (4.7)-(4.8), належить до класу задач ЗЦЛП. Для формального визначення цього класу задач наведемо стандартне означення.

Визначення 3 [128] Нехай задано матрицю обмежень $A \in \mathbb{R}^{m \times d}$, вектор обмежень $b \in \mathbb{R}^m$, вектор коефіцієнтів цільової функції $c \in \mathbb{R}^d$ та підмножину $I \subseteq \{1, \dots, d\}$ індексів цілочисельних змінних. Задача (A, b, c, I) формулюється як:

$$z^* = \min\{c^T(w_j, \dots, \delta_j) \mid A(w_j, \dots, \delta_j) \le b,\ w_j \in \mathbb{R}^d,\ \delta_j \in \mathbb{Z},\ \forall j \in I\} \tag{4.21}$$

Вектори з множини $W = \{(w_j, \delta_j) \mid A(w_j, \dots, \delta_j) \le b,\ w_j \in \mathbb{R}^d,\ \delta_j \in \mathbb{Z},\ \forall j \in I\}$ називаються допустимими розв'язками задачі ЗЦЛП. Допустимий розв'язок $w^* \in W_{\text{MILP}}$ є оптимальним, якщо значення цільової функції задовольняє $c^T w^* = z^*$.

Задача (4.7)-(4.8) містить неперервні транспортні змінні $x_{ij} \geq 0$ та булеві змінні вибору $\delta_i \in \{0,1\}$, що формалізують обмеження кардинальності $\sum_{i=1}^{N} \delta_i \leq K$. Проте ця задача не сформульована у канонічній формі, що унеможливлює пряме застосування стандартних розв'язувачів, зокрема методу гілок та границь [44]. Для зведення до канонічної форми необхідно включити змінні вибору δ_i до цільової функції з нульовими коефіцієнтами, що зберігає значення цільової функції, водночас дозволяючи формалізувати обмеження у матричній формі:

$$\min_{\delta_i, x_{ij}} \sum_{i=1}^{N} \sum_{j=1}^{N} \operatorname{dist}(\xi_i, \xi_j)^r \, x_{ij} + 0 \cdot \delta_i \tag{4.22}$$

У цій формі змінні вибору δ_i не впливають на відстань Канторовича–Рубінштейна, проте їх присутність у цільовій функції дозволяє побудувати єдину матрицю обмежень $A_{(2N+1)\times(N^2+N)}$, що об'єднує всі обмеження задачі. Застосовуючи аналогічний підхід до обмежень (4.7)-(4.8), отримаємо їхній канонічний запис у формі двобічних нерівностей:

$$\begin{array}{ll} \sum_{i=1}^{N} x_{ij} = p_j & \Longrightarrow \; p_j \leq \sum_{i=1}^{N} \sum_{j=1}^{N} x_{ij} - 0 \cdot \delta_j \leq p_j, \\ \sum_{j=1}^{N} x_{ij} \leq \delta_i & \Longrightarrow \; -\infty \leq \sum_{i=1}^{N} \sum_{j=1}^{N} x_{ij} - \delta_i \leq 0, \\ \sum_{i=1}^{N} \delta_i \leq K & \Longrightarrow \; K \leq \sum_{i=1}^{N} \sum_{j=1}^{N} 0 \cdot x_{ij} + \delta_i \leq K. \end{array} \tag{4.23}$$

Перший рядок перетворює обмеження балансу попиту (4.7) на двобічну нерівність з однаковими верхньою та нижньою межами, що еквівалентно рівності. Другий рядок представляє обмеження пропускної спроможності (4.8) з нижньою межею $-\infty$, що відповідає нерівності типу «менше або дорівнює». Третій рядок фіксує кількість активних центроїдів рівно K, перетворюючи обмеження кардинальності (4.8) на рівність. Матриця обмежень $A_{(2N+1)\times(N^2+N)}$ канонічної форми має наступну блочну структуру:

$$A = \left(\begin{array}{cccccccccc|c} 1 & \cdots & 0 & 1 & \cdots & 0 & \cdots & 1 & \cdots & 0 & \\ 0 & \cdots & 1 & 0 & \cdots & 1 & \cdots & 0 & \cdots & 1 & \mathbf{0}_{I\times I} \\ \vdots & & \vdots & \vdots & & \vdots & \ddots & \vdots & & \vdots & \\ \hline 1 & \cdots & 1 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & \\ 0 & \cdots & 0 & 1 & \cdots & 1 & \cdots & 0 & \cdots & 0 & -\mathbf{I}_I \\ \vdots & & \vdots & \vdots & & \vdots & \ddots & \vdots & & \vdots & \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 1 & \cdots & 1 & \\ \hline \multicolumn{9}{c|}{\mathbf{0}_{1\times I^2}} & \mathbf{1}_{1\times I} \end{array}\right) \begin{array}{l} \left.\vphantom{\begin{array}{c}a\\a\\a\end{array}}\right\} I \\ \left.\vphantom{\begin{array}{c}a\\a\\a\end{array}}\right\} I \\ \}1 \end{array} \tag{4.24}$$

Структура матриці (5.29) складається з трьох горизонтальних блоків. Верхній блок розмірності кодує обмеження балансу попиту (4.8): для кожного стовпця сумуються відповідні транспортні змінні за всіма джерелами , при цьому змінні

вибору не беруть участі. Середній блок розмірності кодує обмеження пропускної спроможності (4.8): для кожного рядка сумуються транспортні змінні за всіма призначеннями із від'ємним одиничним блоком для змінних вибору. Нижній рядок кодує обмеження кардинальності (4.8). Відповідні вектори коефіцієнтів цільової функції та двобічних обмежень мають вигляд:

$$\begin{aligned} c &= \left(\underbrace{\mathrm{dist}(\xi_i,\xi_j)^r,\dots,\mathrm{dist}(\xi_i,\xi_j)^r}_{N^2}, \underbrace{0,\dots,0}_{N} \right), \\ b_l &= \left(\underbrace{p_1,\dots,p_N}_{N}, \underbrace{-\infty,\dots,-\infty}_{N}, \underbrace{K}_{1} \right), \\ b_u &= \left(\underbrace{p_1,\dots,p_N}_{N}, \underbrace{0,\dots,0}_{N}, \underbrace{K}_{1} \right). \end{aligned} \tag{4.25}$$

Канонічна форма (4.22)-(4.25) повністю визначає задачу ЗЦЛП з обмеженнями кардинальності, що дозволяє застосувати стандартні алгоритми розв'язання, зокрема метод гілок та границь із площинами відсікання, який описано в наступному підрозділі.

Загальноприйнятим підходом до знаходження оптимального розв'язку задачі ЗЦЛП (4.21) є метод гілок та границь (Branch-and-Bound, B&B) – сімейство алгоритмів, побудованих на розбитті вихідної задачі на підзадачі з різними допустимими областями, що утворюють дерево пошуку [44,109]. Нехай $\mathcal{D}$ позначає простір пошуку, що містить допустимі розв'язки вихідної задачі (4.21), з цільовою функцією $f : \mathcal{D} \to \mathbb{R}$: $\mathcal{P} = (\mathcal{D}, f)$.

На кожній ітерації алгоритм B&B обирає нову підмножину $\mathcal{S} \subset \mathcal{D}$ простору пошуку для дослідження з черги $\mathcal{L}$ недосліджених підмножин. Якщо розв'язок $\hat{w}' \in \mathcal{S}$ забезпечує краще значення цільової функції порівняно з поточним найкращим розв'язком $\hat{w}$, тобто $f(\hat{w}') < f(\hat{w})$, глобальний розв'язок оновлюється. Якщо жоден розв'язок у $\mathcal{S}$ не може покращити поточний найкращий, тобто $f(\hat{w}) \le f(w)$ для всіх $w \in \mathcal{S}$, підмножина відсікається. В іншому випадку підмножина $\mathcal{S}$ розбивається на дочірні підзадачі $\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_r$, які додаються до черги $\mathcal{L}$. Алгоритм завершується, коли черга $\mathcal{L}$ стає порожньою:

1. Ініціалізація: покласти $\mathcal{L} = \{\mathcal{D}\}$ та ініціалізувати $\hat{w}$;
2. Пошук: обрати підзадачу $\mathcal{S}$ з $\mathcal{L}$ для дослідження. Якщо існує розв'язок $\hat{w}' \in \{w \in \mathcal{S} \mid f(w) < f(\hat{w})\}$, покласти $\hat{w} = \hat{w}'$;

3. Оцінювання: якщо $\mathcal{S}$ не може бути відсічена, розбити її на підзадачі $\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_r$ та додати їх до $\mathcal{L}$;
4. Видалення: видалити $\mathcal{S}$ з $\mathcal{L}$.

Метод B&B допускає множину стратегій побудови дерева пошуку, включаючи евристики пошуку [22], стратегії вибору змінних для розгалуження [69] тощо. Ключовим компонентом методу є лінійна релаксація, що полягає у знятті обмеження цілочисловості для змінних вибору δ_i у кожній підзадачі дерева B&B:

$$\hat{z} = \min\{c^T(w_j, \dots, \delta_j) \mid A(w_j, \dots, \delta_j) \leq b,\ w_j, \delta_j \in \mathbb{R}^d\} \tag{4.26}$$

Якщо оптимальне значення лінійної релаксації (4.26) є не меншим за поточний найкращий розв'язок z^*, відповідний вузол дерева може бути відсічений без подальшого дослідження [44]. Алгоритми розв'язання задач лінійного програмування мають широке висвітлення у літературі [23,51,115].

Для підвищення ефективності відсікання вузлів дерева пошуку ми застосовуємо метод площин відсікання [37], який генерує додаткові обмеження, що звужують допустиму область лінійної релаксації шляхом виключення дробових розв'язків при збереженні всіх цілочислово-допустимих точок. Розглянемо вузол дерева з допустимою областю $\mathcal{S}$, визначеною багатогранником $P = \{w \in \mathbb{R}^d_+ : Aw = b\}$, де шукаються цілочислові розв'язки $W = \{w \in \mathbb{Z}^d_+ : Aw = b\}$ з цілочисловими A та b. Нехай w^* є оптимальним розв'язком лінійної релаксації $\min\{c^T w : w \in P\}$ у поточному вузлі, отриманим за базисом $B \subseteq \{1, \dots, d\}$ матриці A з $w^*_B = A_B^{-1}b - A_B^{-1}A_{\bar{B}}w_{\bar{B}}$ та $w^*_{\bar{B}} = 0$, де $\bar{B} = \{1, \dots, d\} \setminus B$.

Якщо w^* є цілочисельними, поточний вузол надає допустимий розв'язок для оцінювання на кроці пошуку B&B. В іншому випадку метод генерує валідну нерівність, що відсікає w^* від допустимої області, зберігаючи при цьому всі цілочисельні точки. Для деякого індексу $i \in B$ з $w^*_i \notin \mathbb{Z}$ кожний допустимий цілочисловий розв'язок $w \in W$ повинен задовольняти [73]:

$$A_{i\cdot}^{-1}b - \sum_{j \in \bar{B}} A_{i\cdot}^{-1}A_{\cdot j}w_j \in \mathbb{Z} \tag{4.27}$$

де $D_{i\cdot}$ позначає i-й рядок матриці D. Застосовуючи функцію дробової частини $f(\alpha) = \alpha - \lfloor\alpha\rfloor$ для $\alpha \in \mathbb{R}$, що зберігає цілочисельність при використанні властивості $0 \leq f(\cdot) < 1 \quad w \geq 0$, отримаємо площину відсікання:

$$\sum_{j\in\bar{B}} f(A_{i\cdot}^{-1}A_{\cdot j})w_j \geq f(A_{i\cdot}^{-1}b) \tag{4.28}$$

яка є валідною для $P_I = \text{conv}(W)$ та порушується розв'язком w^*, оскільки $w^*_{\bar{B}} = 0 \quad f(A_{i\cdot}^{-1}b) = f(w_i^*) > 0$ [37]. Для задачі ЗЦЛП з $W = \{w \in \mathbb{Z}_+^p \times \mathbb{R}_+^{d-p} : Aw = b\}$ підхід узагальнюється через диз'юнктивний аргумент. Використовуючи позначення $\tilde{a}_j = A_{i\cdot}^{-1}A_{\cdot j}$, $\tilde{b} = A_{i\cdot}^{-1}b$, $f_j = f(\tilde{a}_j)$, $f_0 = f(\tilde{b})$ і розбиття $\bar{B}^+ = \{j \in \bar{B} : \tilde{a}_j \geq 0\}$, $\bar{B}^- = \bar{B} \setminus \bar{B}^+$, площина відсікання для задачі ЗЦЛП має вигляд:

$$\sum_{\substack{j:f_j\leq f_0\\ j\in\mathbb{Z}_+^d}} f_j w_j + \sum_{\substack{j:f_j>f_0\\ j\in\mathbb{Z}_+^d}} \frac{f_0(1-f_j)}{1-f_0} w_j + \sum_{\substack{j\in\bar{B}^+\\ j\in\mathbb{R}_+^d}} \tilde{a}_j w_j - \sum_{\substack{j\in\bar{B}^-\\ j\in\mathbb{R}_+^d}} \frac{f_0}{1-f_0}\tilde{a}_j w_j \geq f_0 \tag{4.29}$$

Інтеграція методу площин відсікання у структуру алгоритму B&B відбувається на етапі оцінювання, що посилює алгоритм наступним чином:

1. Ініціалізація: позначити $\mathcal{L} = \{\mathcal{D}\}$ та ініціалізувати $\hat{w}$ допустимим розв'язком або $+\infty$;
2. Вибір вузла: обрати підзадачу $\mathcal{S}$ з $\mathcal{L}$ для дослідження;
3. Лінійна релаксація: розв'язати лінійну релаксацію $\min\{c^T w : w \in P_{\mathcal{S}}\}$ та отримати w^* зі значенням цільової функції z^*;
4. Оцінювання границь: якщо $z^* \geq f(\hat{w})$, відсікти $\mathcal{S}$ та перейти до кроку видалення.
5. Побудова площин відсікання: якщо w^* є дробовим у цілочислових компонентах:
    1. побудувати площину відсікання (4.28) або (4.29), що порушується w^*;
    2. додати площину до $P_{\mathcal{S}}$, формуючи посилену релаксацію $P'_{\mathcal{S}} \subset P_{\mathcal{S}}$;
    3. повернутися до кроку 3 з посиленим формулюванням або перейти до розгалуження, якщо досягнуто ліміт відсікань.
6. Оновлення: якщо $w^* \in W$ та $f(w^*) < f(\hat{w})$, покласти $\hat{w} = w^*$.
7. Розгалуження: якщо $w^* \notin W$ після генерації площин відсікання, обрати дробову змінну w_k та розбити $\mathcal{S}$ на $\mathcal{S}_1 = \mathcal{S} \cap \{w : w_k \leq \lfloor w_k^* \rfloor\}$ та $\mathcal{S}_2 = \mathcal{S} \cap \{w : w_k \geq \lceil w_k^* \rceil\}$, додавши їх до $\mathcal{L}$.
8. Видалення: видалити $\mathcal{S}$ з $\mathcal{L}$.
9. Повторювати кроки 2-8 до $\mathcal{L} = \emptyset$, після чого повернути $\hat{w}$ як оптимальний розв'язок.

Додавання площин відсікання у кожному вузлі посилює нижні оцінки по всьому дереву пошуку, що потенційно дозволяє раніше відсікати підзадачі та зменшувати загальну кількість досліджених вузлів [73]. Площини відсікання, додані у вузлі, залишаються валідними для всіх дочірніх вузлів дерева B&B, оскільки останні успадковують обмеження батьківських вузлів. Важливим результатом є те [6], що площини відсікання, отримані у вузлі, можуть бути посилені для збереження валідності у всьому дереві перебору, а не лише у вузлі та його нащадках, де вони були спочатку обчислені. Ця властивість глобальної валідності забезпечує суттєві переваги для чисельної стабільності та обчислювальної ефективності.

Розв'язання задачі ЗЦЛП з матрицею обмежень $A_{(2N+1)\times(N^2+N)}$ для N елементів може бути обчислювально складним при зростанні розміру вхідної множини. Для подолання цього обмеження ми пропонуємо жадібний алгоритм попереднього відбору $N > \hat{K} > K$ кандидатів для центроїдів, що дозволяє зменшити розмірність задачі без суттєвої втрати точності. Жадібний алгоритм використовує аналогію з центром мас, де ймовірності p_i відповідають «масі» кожного елемента, а алгоритм ітеративно ідентифікує центр кожного кластера. На кожній ітерації алгоритм оновлює «масу» елемента відносно його найближчого сусіда.

1. Ініціалізація: задано $m = \hat{K}$, $N \geq m$, $\{\xi_i,\, i = 1, \dots, N\}$, $\{p_i,\, i = 1, \dots, N\}$. Покласти $k = 0$, $m^k = N$, $W^k = 0$, $\xi_i^k = \xi_i$, $p_i^k = p_i$, $i = 1, \dots, N$.
2. Пошук: знайти i^{k+1}, j^{k+1} такі, що:
$$\|\xi_{i^{k+1}} - \xi_{j^{k+1}}\| = \min_{i,j:\, p_i^k > 0,\, p_j^k > 0} \|\xi_i - \xi_j\| \tag{4.30}$$
3. Оновлення маси: якщо $p_{i^{k+1}}^k \geq p_{j^{k+1}}^k$, покласти $p_{i^{k+1}}^{k+1} = p_{i^{k+1}}^k + p_{j^{k+1}}^k$, $p_{j^{k+1}}^{k+1} = 0$; інакше $p_{i^{k+1}}^{k+1} = 0$, $p_{j^{k+1}}^{k+1} = p_{j^{k+1}}^k + p_{i^{k+1}}^k$.
4. Збереження: $p_i^{k+1} = p_i^k$ для всіх $i \neq i^{k+1}, j^{k+1}$.
5. Оновлення відстані: $W^{k+1} = W^k + p_{i^{k+1}}^k \cdot \|\xi_{i^{k+1}} - \xi_{j^{k+1}}\|$, $m^{k+1} = m^k - 1$.
6. Якщо $m^{k+1} > m$, позначити $k := k + 1$ та повторити кроки 2-5. Інакше визначити центри як $\bigcup_{i:\, p_i^{k+1} > 0} \xi_i$ та транспортну відстань $W = W^{k+1}$.

Інтуїтивна інтерпретація алгоритму (4.30) полягає в наступному: на кожній ітерації знаходиться пара найближчих елементів із додатними масами, після чого маса елемента з меншою вагою передається елементу з більшою вагою, а сам елемент прибирається з множини. Цей процес ітеративно зменшує кількість активних

елементів від N до $\hat{K}$, причому елементи, що накопичили найбільшу масу, є кандидатами на роль центроїдів, оскільки вони розташовані у найбільш щільних ділянках розподілу.

Параметр $\hat{K}$ обирається емпірично. Хоча жадібний метод може бути застосований самостійно як метод ініціалізації, він не досягає тієї ж точності, що й методи (4.20) та (4.21). Алгоритм є чутливим до викидів, тому його основне призначення полягає у відборі початкових $\{\xi_i\}_{i=1}^{\hat{K}}$ кандидатів для подальшого використання у задачі ЗЦЛП (4.7)-(4.8), де з $\hat{K}$ кандидатів обираються оптимальні K центроїдів. Таким чином, комбінований підхід «жадібний алгоритм + ЗЦЛП» зменшує розмірність матриці обмежень з $A_{(2N+1)\times(N^2+N)}$ до $A_{(2\hat{K}+1)\times(\hat{K}^2+\hat{K})}$, що суттєво прискорює розв'язання при збереженні якості ініціалізації.

Як зазначено в огляді методів ЗЦЛП [44], ефективні стратегії розгалуження повинні забезпечувати баланс між якістю прийнятих рішень та часом, необхідним для прийняття кожного рішення. На практиці швидкість знаходження розв'язку залежить від кількох факторів, включаючи щільність лінійної релаксації, розмір задачі, діапазон коефіцієнтів та розрідженість задачі, що впливає на побудову графового представлення та обчислення ознак. Запропонований метод з обмеженням кардинальності було оцінено на кількох відкритих наборах даних із числовими ознаками, перелічених у таблиці 4.1, для оцінювання обчислювальної ефективності та точності порівняно з існуючим підходом (4.20).

Табл. 4.1: Відкриті набори даних для порівняльного аналізу ефективності запропонованого методу.

Набір даних	Класи	Загальна кількість	Розмірність
Іриси Фішера [34]	3	150	4
Розпізнавання вин [116]	3	178	13
Рукописні цифри [30]	10	5620	64

Вихідний код та результати експериментів є загальнодоступними у репозиторії GitHub [58]. В якості програмного забезпечення розв’язання ЗЦЛП обрано SCIP [2]

з інтеграцією Google OR-Tools [97], оскільки на момент проведення дослідження він є одним із найшвидших некомерційних розв'язувачів для задач ЗЦЛП з підтримкою багатопотокових обчислювальних середовищ. Кожний алгоритм оцінювався на 16-ядерному процесорі; результати порівняння часу знаходження оптимального розв'язку наведено у таблиці 4.2, а відстані Канторовича–Рубінштейна (точності) – у таблиці 4.3. Основною метою експериментів є визначення оптимального емпіричного значення параметра $\hat{K}$ у комбінованому жадібному (4.30) та алгоритмі ЗЦЛП (4.21), що забезпечує кращу точність та швидкість пошуку порівняно з K-середніх++ (4.20). Точність K-середніх++ оцінювалась шляхом усереднення відстані Канторовича–Рубінштейна за кількістю експериментальних запусків, зазначених у таблиці 5.3, тоді як комбінований метод оцінювався шляхом фіксації кількості потенційних кандидатів $\hat{K}$, обраних жадібним алгоритмом (4.30), з подальшим відбором K центроїдів алгоритмом ЗЦЛП (4.21). Аналіз результатів, представлених у таблицях 4.2 та 4.3, виявляє характерний компроміс між обчислювальною вартістю та якістю ініціалізації. Алгоритм K-середніх++ демонструє малий час виконання, що практично не залежить від кількості запусків для наборів даних малої та середньої розмірності.

Табл. 4.2: Порівняння часу виконання алгоритмів ініціалізації.

Набір даних		Іриси Фішера	Розпізнавання вин	Рукописні цифри
K-середніх++ (кількість експериментів)	100	0.0001с	0.0002с	0.0010с
	1000	0.0001с	0.0001с	0.0011с
	10000	0.0001с	0.0001с	0.0010с
Жадібний + ЗЦЛП (кількість кандидатів на розв’язок)	5	0.05с	0.11с	221.47с
	50	0.69с	0.6с	229.15с
	150	13.83с	12.34с	250.67с

Табл. 4.3: Порівняння відстані Канторовича–Рубінштейна алгоритмів ініціалізації.

Набір даних		Іриси Фішера	Розпізнавання вин	Рукописні цифри
K-середніх++ (кількість експериментів)	100	156.69	23101.47	74866.93
	1000	150.46	23217.78	74724.67
	10000	150.23	23272.16	74744.20
Жадібний + ЗЦЛП (кількість кандидатів на розв’язок)	5	133.39	20645.25	77093.97
	50	108.4	19490.35	69470.51
	150	98.13	16376.97	68840.82

Проте відстань Канторовича–Рубінштейна, усереднена за запусками, суттєво перевищує значення, досягнуті комбінованим методом. Зокрема, для набору даних «Іриси Фішера» комбінований підхід із $\hat{K} = 150$ забезпечує відстань 98.13, що становить лише 65% від найкращого результату K-середніх++ (150.23 при 10000 запусках). Для набору «Розпізнавання вин» покращення є ще більш суттєвим: комбінований метод досягає відстані 16376.97 порівняно з 23101.47 для K-середніх++, що відповідає зменшенню на 29%.

Для набору даних «Рукописні цифри» з високою розмірністю ($d = 64$) та великою кількістю елементів ($N = 5620$) спостерігається очікуване зростання обчислювальної вартості комбінованого методу до 221-251 секунд, проте при $\hat{K} \geq 50$ досягається суттєве покращення точності порівняно з K-середніх++. Зростання часу виконання обумовлене розміром матриці обмежень (4.24): для $\hat{K} = 150$ матриця має розмірність 301×22650, що є обчислювально складною задачею для методу гілок та границь. Візуальне порівняння точності розв’язку зображене на рисунку 4.1.

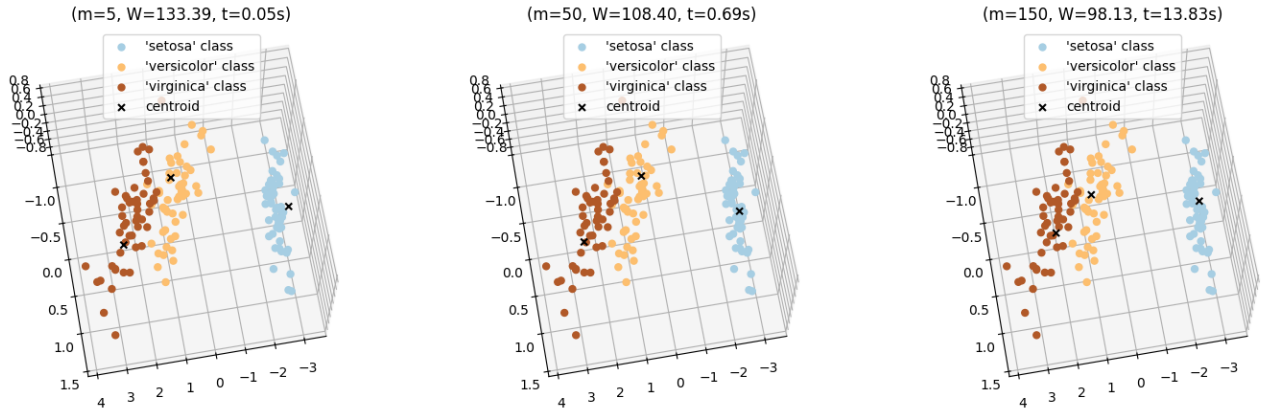

Рис. 4.1: Візуалізація пошуку комбінованого жадібного та алгоритму ЗЦЛП з 3D-проекцією набору даних.

Підсумовуючи, результати експериментів дозволяють зробити три ключові висновки щодо стратегії ініціалізації центроїдів:

1. комбінований підхід «жадібний алгоритм + ЗЦЛП» забезпечує кращу точність ініціалізації, виміряну відстанню Канторовича–Рубінштейна, порівняно з K-середніх++ навіть при великій кількості стохастичних запусків останнього;
2. параметр $\hat{K}$ виконує роль регулятора компромісу між обчислювальною вартістю та якістю ініціалізації: збільшення $\hat{K}$ монотонно покращує точність за рахунок зростання часу розв'язання задачі ЗЦЛП;
3. для задач із помірною кількістю елементів ($N \leq 10^3$) комбінований підхід забезпечує прийнятний час виконання при суттєвому покращенні якості ініціалізації, що робить його практично придатним для побудови початкового наближення у задачі стохастичної квазіградієнтної кластеризації (4.3)-(4.4).

Властивість детермінованості запропонованого підходу є його перевагою, оскільки на відміну від стохастичного K-середніх++ (4.20), комбінований метод забезпечує відтворюваність результатів та гарантоване зменшення цільової функції при збільшенні параметра $\hat{K}$, що є суттєвим для застосувань, де вимагається надійна та передбачувана якість ініціалізації.

4.4 Збіжність методу стохастичної квазіградієнтної кластеризації

В попередньому розділі, присвяченому огляду стохастичних методів оптимізації (2.6), (2.13), (2.19), а також їх модифікацій з адаптивним кроком (2.48), (2.52), (2.58) було показано за яких умов та з якою швидкістю дані методи оптимізації збігаються до точки оптимуму. Оскільки цільова функція (4.13) є узагальнено-диференційованою, для отримання оцінки збіжності для методу SQC, необхідно застосувати теорему 16 про збіжність ітераційного методу узагальненого градієнта.

Теорема 18 (Збіжність методу стохастичної квазіградієнтної кластеризації) Нехай існує послідовність $\{y^t = (y_1^t, \dots, y_K^t)\}$, отримана рекурентним співвідношенням $y_k^{t+1} = \Pi_Y(y_k^t - \rho_t g_k(\tilde{\xi}^t))$, $\Pi_Y(\cdot) = \arg\min_{y \in Y} \| \cdot - y \|$, підмножина елементів $\{\tilde{\xi}^t, t = 0, 1, \dots\}$ з $\{\xi_i\}$ та послідовність $\{\rho_t\}$:

$$\rho_t > 0, \sum_{t=0}^{\infty} \rho_t = \infty, \sum_{t=0}^{\infty} \rho_t^2 < \infty \tag{4.31}$$

Позначимо через $F(Y^*)$ множину значень F в оптимальних точках Y^* задачі, де $Y^* = \{y = (y_1, \dots, y_K) : \partial F(y) \in N_Y(y_1) \times \dots \times N_Y(y_K)\}$ і $N_Y(y_k)$ позначає нормальний конус до множини Y у точці y_k. Якщо $F(Y^*)$ не містить проміжків і множина Y є опуклим компактом, то з ймовірністю один $\{y^t\}$ збігається до зв'язної компоненти Y^*, а послідовність $\{F(y^t)\}$ має границю.

Доведення. Для застосування теореми 16 необхідно встановити узагальнену диференційовність цільової функції F, незміщеність та обмеженість стохастичної оцінки субградієнта $g(\tilde{\xi}^t)$, а також локальну мінімізуючу властивість стохастичної проекційної ітерації. Розпочнемо з аналізу структури цільової функції F, заданої виразом (4.13). Для кожного $i \in \{1, \dots, N\}$ $k \in \{1, \dots, K\}$ функція $\psi_{ik}(y_k) = \|y_k - \xi_i\|^r$ є неперервно диференційованою при $r \geq 2$ із градієнтом $\nabla \psi_{ik}(y_k) = r\|y_k - \xi_i\|^{r-2}(y_k - \xi_i)$ і, отже, узагальнено-диференційованою з $G_{\psi_{ik}}(y_k) = \{\nabla \psi_{ik}(y_k)\}$ (теорема 10). При $r \in [1, 2)$ функція ψ_{ik} є опуклою як композиція неперервно диференційовної монотонно зростаючої функції $u \mapsto u^r$ на $[0, \infty)$ та опуклої норми, отже за теоремою 11 вона також узагальнено-диференційовна із псевдоградієнтом, що збігається з її субдиференціалом у сенсі опуклого аналізу. Застосовуючи теорему 15 про субградієнт функції мінімуму до $\varphi_i(y) = \min_{1 \leq k \leq K} \psi_{ik}(y_k)$, отримуємо узагальнену диференційовність φ_i із псевдоградієнтним відображенням

$$G_{\varphi_i}(y) = \operatorname{co}\{(0_d, \dots, g_{k^*}, \dots, 0_d) : k^* \in S(\xi_i, y),\ g_{k^*} \in G_{\psi_{ik^*}}(y_{k^*})\} \tag{4.32}$$

де ненульова компонента g_{k^*} розташована на k^*-й позиції. Цільова функція є невід'ємною скінченною лінійною комбінацією $F(y) = \sum_{i=1}^{N} p_i \varphi_i(y)$. За правилом ланцюжка для узагальнено-диференційовних функцій (теорема 12), застосованим до зовнішньої функції $f_0(z_1, \dots, z_N) = \sum_{i=1}^{N} p_i z_i$, отримуємо узагальнену диференційовність F із псевдоградієнтним відображенням:

$$G_F(y) = \operatorname{co}\Big\{\sum_{i=1}^{N} p_i g^{(i)} : g^{(i)} \in G_{\varphi_i}(y),\ i = 1, \dots, N\Big\} \tag{4.33}$$

Оскільки G_F є локально обмеженим напівнеперервним зверху відображенням, а Y^K – компактом, F є ліпшицевою на Y^K зі сталою $L_F < \infty$ (теорема 10) та існує константа $\Gamma < \infty$ така, що:

$$\sup_{y \in Y^K} \sup_{g \in G_F(y)} \|g\| \le \Gamma \tag{4.34}$$

Уведемо ймовірнісний простір $(\Omega, \mathcal{F}, \mathbb{P})$, на якому визначено незалежні випадкові елементи $\{\tilde{\xi}^t\}_{t \ge 0}$ з розподілом $\mathbb{P}(\tilde{\xi}^t = \xi_i) = p_i$, та функцією відбору $\mathcal{F}_t = \sigma(\tilde{\xi}^0, \dots, \tilde{\xi}^{t-1})$, відносно якої y^t є $\mathcal{F}_t$-вимірним. Зафіксуємо детерміністичне борелеве правило $\tilde{k}^*(\xi, y) \in S(\xi, y)$ вибору індекса найближчого центру при наявності зв'язок (наприклад, найменший індекс із аргмінімуму) і означимо стохастичну оцінку згідно з (4.16):

$$g_k(\tilde{\xi}^t, y^t) = \begin{cases} r\|\tilde{\xi}^t - y_k^t\|^{r-2}(y_k^t - \tilde{\xi}^t), & k = \tilde{k}^*(\tilde{\xi}^t, y^t), \\ 0_d, & k \ne \tilde{k}^*(\tilde{\xi}^t, y^t). \end{cases} \tag{4.35}$$

Обчислюючи умовне математичне сподівання за функцією відбору $\mathcal{F}_t$, отримуємо

$$\bar{g}(y^t) := \mathbb{E}[g(\tilde{\xi}^t, y^t) \mid \mathcal{F}_t] = \sum_{i=1}^{N} p_i\, g(\xi_i, y^t) \tag{4.36}$$

причому кожен доданок $g(\xi_i, y^t)$ за конструкцією (4.35) належить до множини (4.32) для φ_i (при $r \in [1, 2)$ значення $\xi_i = y_{k^*}^t$ можуть породжувати множинний субградієнт, але утворюють подію нульової міри відносно неперервної динаміки $\{y_k^t\}$). Отже, $\bar{g}(y^t)$ є борелевою однозначною секцією псевдоградієнтного відображення $G_F(y^t)$, тобто $\bar{g}(y^t) \in G_F(y^t)$. Із компактності Y та скінченності множини $\{\xi_i\}$ випливає скінченність константи:

$$D = \max_{1\le i\le N} \sup_{y_k\in Y} \|y_k - \xi_i\| < \infty \tag{4.37}$$

тому стохастична оцінка задовольняє рівномірну оцінку

$$\|g(\tilde\xi^t, y^t)\| \le rD^{r-1} =: M, \qquad \mathbb{E}[\|g(\tilde\xi^t, y^t)\|^2 \mid \mathcal{F}_t] \le M^2 \tag{4.38}$$

Оператор проекції Π_Y на опуклий компакт Y є нерозширювальним, тому для будь-якого $y^* \in Y^K$ покомпонентне застосування властивості $\|\Pi_Y(z) - y_k^*\| \le \|z - y_k^*\|$ до оновлення $y_k^{t+1} = \Pi_Y(y_k^t - \rho_t g_k(\tilde\xi^t, y^t))$ та підсумовування по $k = 1, \dots, K$ дає

$$\|y^{t+1} - y^*\|^2 \le \|y^t - y^*\|^2 - 2\rho_t\langle g(\tilde\xi^t, y^t), y^t - y^*\rangle + \rho_t^2\|g(\tilde\xi^t, y^t)\|^2 \tag{4.39}$$

Із обмеженості (4.38) та умови $\sum_t \rho_t^2 < \infty$ безпосередньо випливає сходження послідовних різниць до нуля з імовірністю один:

$$\|y^{t+1} - y^t\| \le \rho_t M \to 0, \qquad \sum_{t=0}^{\infty}\|y^{t+1} - y^t\|^2 \le M^2\sum_{t=0}^{\infty}\rho_t^2 < \infty \tag{4.40}$$

Обчислюючи умовне математичне сподівання обох частин (4.39) за функцією відбору $\mathcal{F}_t$ та використовуючи незміщеність (4.36), отримуємо стохастичну дисипативну нерівність

$$\mathbb{E}[\|y^{t+1} - y^*\|^2 \mid \mathcal{F}_t] \le \|y^t - y^*\|^2 - 2\rho_t\langle \bar g(y^t), y^t - y^*\rangle + \rho_t^2 M^2 \tag{4.41}$$

що є аналогом нерівності (3.32), отриманої при доведенні теореми 17 для згладженої функції. Розклавши стохастичну оцінку як $g(\tilde\xi^t, y^t) = \bar g(y^t) + \nu^t$ із мартингальним приростом $\nu^t = g(\tilde\xi^t, y^t) - \bar g(y^t)$, $\mathbb{E}[\nu^t \mid \mathcal{F}_t] = 0$, $\mathbb{E}[\|\nu^t\|^2 \mid \mathcal{F}_t] \le M^2$, отримуємо для будь-якої $\mathcal{F}_t$-вимірної обмеженої випадкової величини u^t адаптовану до фільтрації послідовність випадкових величин $\varphi^t = \rho_t\langle \nu^t, u^t\rangle$, що задовольняє

$$\mathbb{E}[\varphi^t \mid \mathcal{F}_t] = 0, \qquad \sum_{t=0}^{\infty}\mathbb{E}[(\varphi^t)^2 \mid \mathcal{F}_t] \le 4M^2\operatorname{diam}(Y^K)^2\sum_{t=0}^{\infty}\rho_t^2 < \infty \tag{4.42}$$

тому за лемою про збіжність рядів ([76], лема 4.1) ряд $\sum_t \varphi^t$ збігається з імовірністю один. Зокрема, для будь-якої граничної точки $y^* \in Y^K$ ряд $\sum_t \rho_t\langle \nu^t, y^t - y^*\rangle$ є збіжним.

Перейдемо до встановлення локальної мінімізуючої властивості, що повторює структуру леми 5. Нехай $\{y^{t_s}\}_{s\ge 0}$ – підпослідовність, що збігається до точки $y \notin Y^*$. За означенням Y^*, для будь-якого $g \in G_F(y)$ існує принаймні один компонентний індекс k_0 такий, що $-g_{k_0} \notin N_Y(y_{k_0})$, тобто $\Pi_Y(y_{k_0} - \rho g_{k_0}) \ne y_{k_0}$ при всіх

достатньо малих $\rho > 0$. Внаслідок напівнеперервності зверху псевдоградієнтного відображення G_F, неперервності проекції та невиродженості точки y існує окіл U_y точки y, константа $\eta > 0$ та $\bar{\rho} > 0$ такі, що для всіх $y' \in U_y$, $\rho \in (0, \bar{\rho}]$ виконується наступна оцінка

$$\langle \bar{g}(y'), y' - \Pi_{Y^K}(y' - \rho \bar{g}(y')) \rangle \geq \eta\rho \tag{4.43}$$

де $\Pi_{Y^K} = \Pi_Y \times \cdots \times \Pi_Y$ діє покомпонентно. За формулою скінченних приростів для узагальнено-диференційовних функцій [88] існує $\theta_t \in [0,1]$ та $\tilde{g}^t \in G_F(y^t + \theta_t(y^{t+1} - y^t))$ такі, що $F(y^{t+1}) - F(y^t) = \langle \tilde{g}^t, y^{t+1} - y^t \rangle$. Поєднуючи це із напівнеперервністю G_F, властивістю проекції $\langle y^t - y^{t+1}, y^{t+1} - z \rangle \geq 0$ для $z = y^t - \rho_t g(\tilde{\xi}^t, y^t)$ та оцінкою (4.43), отримуємо при $y^t \in U_y$ та достатньо малих ρ_t:

$$F(y^{t+1}) \leq F(y^t) - \frac{\eta}{2}\rho_t + \rho_t \langle \bar{g}(y^t) - g(\tilde{\xi}^t, y^t), \tilde{u}^t \rangle + C\rho_t^2 \tag{4.44}$$

де $\tilde{u}^t$ є $\mathcal{F}_t$-вимірною обмеженою величиною, що описує напрямок дисипативності (із нормою, не більшою за $\operatorname{diam}(Y^K)$), а константа C залежить від M, Γ та L_F. Введемо моменти виходу із ε-окола: позначимо $k_s = t_s$ та $n_s = \min\{t > k_s : \|y^t - y\| > \varepsilon\}$. Для $\varepsilon < \operatorname{dist}(U_y^c, y)$ та достатньо великих s, нерівність (4.44) застосовується для всіх $t \in [k_s, n_s)$, і її підсумовування дає

$$F(y^{n_s}) \leq F(y^{k_s}) - \frac{\eta}{2} \sum_{t=k_s}^{n_s-1} \rho_t + (S_{n_s} - S_{k_s}) + \tilde{C} \sum_{t=k_s}^{n_s-1} \rho_t^2 \tag{4.45}$$

де $S_t = \sum_{s<t} \rho_s \langle \bar{g}(y^s) - g(\tilde{\xi}^s, y^s), \tilde{u}^s \rangle$ є рядом, що збігається м.н. за (4.42). Із оцінки (4.40) випливає, що для виходу послідовності $\{y^t\}$ з $\varepsilon/2$-окола y необхідно $\sum_{t=k_s}^{n_s-1} \rho_t \geq \varepsilon/(2M)$. Припустивши протилежне, тобто що при певному $\omega \in \Omega$ підпослідовність $\{y^{t_s}\}$ ніколи не виходить за межі ε-окола нестаціонарної точки y для $s \geq s_0$, з нерівності (4.44) та $\sum_t \rho_t = \infty$ при $\sum_t \rho_t^2 < \infty$ отримаємо $F(y^{n_s}) \to -\infty$, що суперечить обмеженості F на компакті Y^K. Отже, послідовність обов'язково виходить з ε-окола з імовірністю один, і тоді (4.45) при граничному переході по s дає

$$\limsup_{s \to \infty} F(y^{n_s}) \leq F(y) - \frac{\eta\varepsilon}{4M} < F(y) \tag{4.46}$$

що є стохастичним аналогом локальної мінімізуючої властивості (3.19) леми 5.

Залишилося застосувати узагальнену теорему про збіжність ітераційного методу узагальненого градієнта (теорема 16) із $W(y) = F(y)$ як функцією Ляпунова

та $X^* = Y^*$ як множиною стаціонарних точок. Послідовність $\{y^t\}$ обмежена через включення $y^t \in Y^K$; сходження послідовних різниць $\|y^{t+1} - y^t\| \to 0$ м.н. забезпечується оцінкою (4.40); локальна мінімізуюча властивість виконується м.н. внаслідок (4.46); відсутність проміжків у множині $F(Y^*)$ є припущенням теореми. Отже, висновок теореми 16 безпосередньо переноситься на стохастичний процес $\{y^t\}$: з імовірністю один усі граничні точки послідовності $\{y^t\}$ належать зв'язній компоненті множини Y^*, а числова послідовність $\{F(y^t)\}$ має скінченну границю. ■

Зауважимо, що умова $F(Y^*)$ є властивою для дискретних задач квантизації з фіксованою скінченною множиною елементів $\{\xi_1, \dots, \xi_N\}$, оскільки множина оптимальних значень функції (4.13) є скінченним або зліченним об'єднанням значень, що відповідають різним розбиттям Вороного, кожне з яких породжує скінченне сімейство ізольованих стаціонарних точок [64]. Адаптивні модифікації SQC-Adam успадковують встановлені властивості за умови обмеженості ефективних кроків $\rho_t / \sqrt{v_{k^*}^t + \varepsilon}$, оскільки покоординатне коригування є еквівалентною заміною послідовності $\{\rho_t\}$ на послідовність $\{\tilde{\rho}_t\}$ зі збереженням умов Роббінса-Монро в асимптотиці [76].

4.5 Застосування методів глобальної оптимізації до цільової функції

Відмінність методу SQC від класичних підходів до кластеризації полягає в наявності явно сформульованої цільової функції (4.13), яка визначає відстань Канторовича–Рубінштейна між фіксованим розподілом Ξ та апроксимуючим розподілом рухомих центрів Y. У методах K-середніх та GMM цільова функція в загальному вигляді не доступна явно: ці методи реалізуються через EM-подібну схему почергових оновлень, де якість локального оптимуму суттєво залежить від випадкової ініціалізації, а значення цільової функції оцінюється лише опосередковано через приписування елементів до центрів та оновлення параметрів моделі. Як наслідок, навіть з теоретично обґрунтованими стратегіями ініціалізації, такими як K-середніх++ (4.20) з $O(\log K)$-апроксимацією у математичному сподіванні [5], окремий запуск EM-алгоритму не гарантує наближення до глобального оптимуму, а потребує усереднення за множиною стохастичних повторень. Натомість стохастична апроксимація (4.12)-(4.13), отримана у

попередньому підрозділі, дозволяє безпосередньо обчислювати значення $F(y)$ для довільної конфігурації центрів $y \in Y^K \subset \mathbb{R}^{dK}$, що ставить задачу SQC у клас задач прямої безумовної оптимізації за скінченновимірною векторною змінною.

Прямий доступ до значень неопуклої негладкої функції $F(y)$ відкриває можливість застосування методів глобальної оптимізації, які не покладаються лише на локальну градієнтну інформацію. Зокрема, теорема 18 стверджує лише збіжність ітерацій SQC до зв'язної компоненти множини стаціонарних точок Y^*, а не до глобального оптимуму, що зумовлено неопуклою структурою F та залежністю результату від початкового наближення y^0. Серед методів глобальної оптимізації, придатних для задач без вимог гладкості та опуклості, є метаевристичні методи на основі популяції, які підтримують одночасно множину кандидатів-розв'язків та забезпечують стохастичне дослідження простору пошуку через взаємодію індивідів. Дані методи мають здатність уникати застрягання у локальних мінімумах завдяки обміну інформацією між індивідами, розташованими у структурно різних областях простору Y^K.

Серед популяційних методів глобальної оптимізації найбільш придатним для задачі SQC є алгоритм DE [118] як простий та стійкий метод мінімізації неопуклих неперервних функцій. Перевага DE над іншими популяційними методами (метод рою частинок [52], метод бджолиних колоній [50], метод мурашиних колоній [26]) у контексті задачі SQC обумовлена двома структурними особливостями. По-перше, DE працює безпосередньо у неперервному просторі $\mathbb{R}^{dK}$, що відповідає умовам параметризації задачі квантизації, без необхідності дискретизації простору пошуку або спеціальних адаптацій, як у мурашиному алгоритмі. По-друге, схема генерації нових кандидатів у DE не використовує «ведучого» елемента (на відміну від глобально найкращої частинки в методі рою частинок [52]), а спирається на стохастичну різницю випадково обраних членів популяції, що автоматично адаптує масштаб збурень до поточного розподілу популяції та робить метод сумісним із гібридизацією, у якій локальні методи оптимізації (зокрема SQC та SQC-Adam) застосовуються як оператори уточнення для окремих індивідів популяції. Ця сумісність показує, що запропонований метод SQC може бути доповненим з DE як метод локальної оптимізації, що зменшує значення цільової функції F для індивіда без зміни його позиції у популяції відносно решти кандидатів.

Сформулюємо повний алгоритм SQC-DE, що поєднує глобальну схему DE зі стохастичною квазіградієнтною локальною оптимізацією. Задано множину елементів $\{\xi_i,\, i = 1, \dots, N\}$ з ймовірностями $\{p_i\}$, кількість центрів K, та гіперпараметри: розмір популяції μ, коефіцієнт масштабування $F_{\mathrm{DE}} \in (0, 2]$, ймовірність кросоверу $\mathrm{CR} \in [0, 1]$, кількість поколінь $G_{\max}$, крок локальної оптимізації $\rho > 0$, кількість локальних ітерацій T_{loc} та параметр згладжування $r = 3$.

1. Ініціалізація популяції. Згенерувати μ випадкових конфігурацій центрів $\{y^{(i),0} = (y_1^{(i),0}, \dots, y_K^{(i),0}) \in Y^K\}_{i=1}^{\mu}$ із застосуванням комбінованого жадібного алгоритму та ЗЦЛП (4.7)-(4.8) або стратегії K-середніх++ (4.20) для кожного індивіда. Обчислити $F(y^{(i),0})$ для $i = 1, \dots, \mu$.
2. Для $G = 0, 1, \dots, G_{\max} - 1$:
   i. Мутація. Для кожного цільового вектора $y^{(i),G}$ обрати три різних випадкових індекси $r_1, r_2, r_3 \in \{1, \dots, \mu\} \setminus \{i\}$ та сформувати мутаційний вектор:
   $$v^{(i),G} = y^{(r_1),G} + F_{\mathrm{DE}} \cdot (y^{(r_2),G} - y^{(r_3),G}) \in \mathbb{R}^{dK} \tag{4.47}$$
   ii. Кросовер. Сформувати пробний вектор $u^{(i),G}$ покоординатним змішуванням:
   $$u_j^{(i),G} = \begin{cases} v_j^{(i),G}, & \mathrm{rand}_j \le \mathrm{CR}, \\ y_j^{(i),G}, & \end{cases} \qquad j = 1, \dots, dK, \tag{4.48}$$
   де $\mathrm{rand}_j \sim \mathrm{Uniform}(0, 1)$, а j_{rand} – випадково обраний індекс координати;
   iii. Локальне уточнення. Застосувати T_{loc} ітерацій алгоритму SQC (або SQC-Adam) із початковим наближенням $u^{(i),G}$, отримавши уточнений пробний вектор $\tilde{u}^{(i),G} = \mathrm{SQC}(u^{(i),G}, T_{\mathrm{loc}}, \rho, r)$;
   iv. Селекція. Оновити популяцію за жадібним критерієм:
   $$y^{(i),G+1} = \begin{cases} \tilde{u}^{(i),G}, & F(\tilde{u}^{(i),G}) \le F(y^{(i),G}), \\ y^{(i),G}, & F(\tilde{u}^{(i),G}) > F(y^{(i),G}). \end{cases} \tag{4.49}$$
3. Повернення найкращого індивіда: $y^* = \arg\min_{1 \le i \le \mu} F(y^{(i),G_{\max}})$.

Ключовою рисою запропонованого алгоритму SQC-DE є двоетапна декомпозиція оптимізаційної задачі: глобальний оператор мутації DE на кроці (i) використовує диференційні збурення $F_{\mathrm{DE}}(y^{(r_2),G} - y^{(r_3),G})$ для дослідження простору Y^K зі стохастичним покриттям областей, недосяжних для локальних методів, тоді як

крок (iii) виконує детерміністичне зменшення F у напрямку, визначеному стохастичною оцінкою субградієнта (4.35), що відповідає теоремі 18 про збіжність SQC до множини стаціонарних точок Y^*. Розмір популяції μ та коефіцієнт масштабування F_{DE} контролюють компроміс між глобальним дослідженням та обчислювальною вартістю однієї ітерації: за рекомендаціями [118], для задач квантизації доцільно обирати $\mu \in [5\,dK, 10\,dK]$, $F_{\mathrm{DE}} = 0.5$ як початковий вибір та $\mathrm{CR} \in \{0.1, 0.9\}$ залежно від бажаного балансу між дослідженням простору та швидкістю локальної концентрації популяції.

Питання теоретичної збіжності гібридного алгоритму SQC-DE до глобального оптимуму функції F розв'язується завдяки результатам, отриманим у роботі [43], де доведено достатні умови глобальної збіжності DE за ймовірністю.

Теорема 19 (Глобальна збіжність DE до оптимального розв’язку) [43] Нехай $\{X(G),\, G = 0, 1, 2, ...\}$ – послідовність популяцій, що генерується алгоритмом DE з елітарною селекцією для розв'язання задачі $\min_{y \in Y^K} F(y)$, де $Y^K \subseteq \mathbb{R}^{dK}$ є метричним простором. Нехай $Y^*_\delta = \{y \in Y^K \mid F(y) - F(y^*) < \delta\}$ є δ-окіл множини глобальних мінімумів із мірою $\mu(Y^*_\delta) > 0$. Якщо існує підпослідовність поколінь $\{G_k\}_{k \geq 1}$ та послідовність додатних чисел $\{\zeta(G_k)\}$ такі, що в популяції $X(G_k)$ існує принаймні один індивід y із відповідним пробним вектором u, для якого

$$\mathbb{P}\{u \in Y^*_\delta\} \geq \zeta(G_k) > 0, \quad \sum_{k=1}^{\infty} \zeta(G_k) = \infty \tag{4.50}$$

то послідовність найкращих знайдених розв'язків збігається за ймовірністю до глобального оптимуму:

$$\lim_{G \to \infty} \mathbb{P}\{X(G) \cap Y^*_\delta \neq \emptyset\} = 1 \tag{4.51}$$

Внеском теореми 19 є те, що збіжність DE до глобального оптимуму не вимагає ергодичності станів популяції, а ґрунтується лише на властивостях нескінченних добутків та елітарної селекції [43]: ймовірність відсутності оптимуму у популяції після G_k поколінь обмежена зверху величиною $\prod_{i=1}^{k-1}(1 - \zeta(i))$, що прямує до нуля за розбіжності ряду $\sum_k \zeta(G_k)$. В роботі [43] також встановлено, що ця умова виконується для модифікованого алгоритму CDE із додатковим оператором допоміжної мутації (рівномірної або гаусової), який гарантує ненульову ймовірність генерації будь-якої точки простору Y^K на кожному поколінні.

Зауважимо, що базовий алгоритм DE без такого допоміжного оператора не задовольняє умовам теореми у загальному випадку, оскільки при потраплянні популяції в окіл локального мінімуму диференційні збурення $y^{(r_2),G} - y^{(r_3),G}$ можуть наближатися до нуля, що порушує оцінку $\mathbb{P}\{u \in Y_\delta^*\} \geq \zeta(G_k) > 0$.

У контексті задачі стохастичної квантизації роль теореми 19 є подвійною. По-перше, вона надає теоретичну гарантію того, що алгоритм SQC-DE при належному виборі оператора допоміжної мутації (наприклад, рівномірне збурення позицій центрів у межах Y) збігається за ймовірністю до глобального мінімуму неопуклої функції (4.13), що неможливо для базового SQC, для якого теорема 18 гарантує лише збіжність до зв'язної компоненти Y^*. По-друге, поєднання теорем 18 та 19 забезпечує структуру збіжності гібридного методу: на рівні окремого індивіда популяції локальна процедура SQC асимптотично досягає стаціонарної точки множини Y^* з імовірністю один, тоді як на рівні всієї популяції диференційна еволюція забезпечує ненульову ймовірність генерації кандидатів у δ-околі глобального мінімуму. Таким чином, запропонований гібридний метод SQC-DE системно подолає залежність якості кластерного розв'язку від випадкової ініціалізації, що є обмеженням EM-подібних схем, та забезпечує теоретично обґрунтовану апроксимацію глобального мінімуму цільової функції стохастичної квазіградієнтної кластеризації.

4.6 Висновки до розділу

Отже, у цьому розділі представлено та обґрунтовано метод SQC як альтернативний підхід до задачі стохастичної квазіградієнтної кластеризації дискретних імовірнісних розподілів. Спочатку задачу квантизації сформульовано як транспортну задачу з рухомими центрами (4.12)-(4.13), що ґрунтується на мінімізації метрики Канторовича–Рубінштейна між фіксованим розподілом елементів та апроксимуючим розподілом рухомих центрів [64]. Показано, що нелінійність цільової функції за позиціями центрів при лінійності за транспортними змінними дозволяє декомпозицію задачі на послідовність підзадач (4.9)-(4.10), кожна з яких є задачею лінійного програмування із гарантованим

монотонним зменшенням відстані між розподілами та оцінкою швидкості збіжності до локального оптимуму $O(\log K)$.

Далі здійснено зведення задачі квантизації до задачі неопуклого стохастичного програмування (4.12)-(4.13) шляхом аналітичного усунення оптимізації за транспортними змінними та ймовірностями центрів, що дозволило побудувати рекурентне правило оновлення позицій центрів (4.16), (4.17) та його адаптивні модифікації SQC-Adam на основі адаптивних градієнтних методів (2.48), (2.52), (2.58). Ключовою перевагою стохастичного підходу є зменшення просторової складності з $O(Nd)$ до $O(d)$, оскільки метод оновлює кожен центроїд лише з використанням одного випадково обраного елемента $\tilde{\xi}$ замість усієї множини вхідних даних $\{\xi_1, \cdots, \xi_N\}$, що робить метод придатним для оброблення великих наборів даних в умовах обмежених обчислювальних ресурсів. Для задачі побудови початкового наближення центроїдів запропоновано детермінований комбінований підхід, що поєднує жадібний алгоритм попереднього відбору кандидатів із задачею ЗЦЛП з обмеженням кардинальності (4.9)-(4.10), та показано його переваги над імовірнісною стратегією K-середніх++ [5] у відтворюваності результатів та гарантованому зменшенні цільової функції при збільшенні параметра $\hat{K}$.

Теоретичну обґрунтованість запропонованого методу підтверджено теоремою 18 про збіжність, доведеною на основі теорії узагальнено-диференційованих функцій [76,88]. Встановлено, що за виконання умов Роббінса–Монро на послідовність кроків спуску $\{\rho_t\}$ та припущення про опуклу компактність допустимої множини Y, послідовність $\{y^t\}$, що генерується алгоритмом SQC, збігається з ймовірністю один до зв'язної компоненти множини стаціонарних точок Y^*, а числова послідовність $\{F(y^t)\}$ має скінченну границю. Адаптивні модифікації SQC-Adam успадковують встановлені властивості збіжності за умови обмеженості ефективних кроків $\rho_t/\sqrt{v_{k^*}^t + \varepsilon}$, що забезпечує застосовність теореми 18 до всієї сім'ї запропонованих стохастичних алгоритмів.

Окремо розглянуто питання глобальної оптимізації цільової функції $F(y)$, що є важливим внаслідок неопуклості задачі квантизації та залежності якості розв'язку від початкового наближення. Запропоновано гібридний алгоритм SQC-DE, що поєднує популяційний метод DE [118] як метод глобальної оптимізації зі

стохастичною квазіградієнтною процедурою SQC як методом локальної оптимізації для уточнення індивідів. На основі теореми 19 [43] показано, що при належному виборі оператора допоміжної мутації гібридний алгоритм SQC-DE збігається за ймовірністю до глобального мінімуму неопуклої функції (4.13), що подолає обмеження класичних EM-подібних схем, чутливих до випадкової ініціалізації.

Запропонований метод стохастичної квантизації не лише усуває обмеження існуючих методів квантизації, пов'язані з масштабованістю та лінійною залежністю просторової складності від кількості елементів вхідної множини, але й забезпечує теоретичні гарантії збіжності, що мотивує його застосування як для класичних задач розміщення об'єктів обслуговування та апроксимації розподілів, так і для сучасних задач стиснення даних у середовищах із обмеженими обчислювальними ресурсами. Прикладне застосування методу SQC та його модифікацій до конкретних задач машинного навчання, зокрема до кольорової квантизації зображень, класифікації високовимірних даних та оновлення індексних структур для наближеного пошуку найближчих сусідів, буде продемонстровано у наступному розділі.

РОЗДІЛ 5. ПРИКЛАДНЕ ЗАСТОСУВАННЯ

У цьому розділі продемонстровано прикладне застосування запропонованого методу стохастичної квантизації SQC та його адаптивних модифікацій до трьох суттєво відмінних класів задач, що ілюструють універсальність розробленого підходу. Спочатку розглянуто задачу кольорової квантизації зображень як окремий випадок задачі апроксимації дискретних розподілів, де метод SQC застосовується для побудови оптимальної колірної палітри з обмеженого набору кольорів. Особлива увага приділена порівнянню запропонованого підходу з класичним алгоритмом K-середніх за обчислювальною складністю та якістю отриманої квантизації на стандартних наборах даних.

Далі досліджено інтеграцію методу SQC з контрастивним навчанням за допомогою архітектури Triplet Network для розв'язання задач класифікації високовимірних даних. Запропонований комбінований підхід долає обмеження методів кластеризації у високовимірних просторах – явище «прокляття розмірності» – шляхом попереднього відображення дискретних даних у низьковимірний латентний простір, у якому евклідова відстань зберігає семантичну подібність елементів. Експериментальна оцінка точності кластеризації виконана на наборі даних MNIST із застосуванням індексу Ранда як метрики якості.

Зрештою, представлено застосування методу SQC для ітеративного оновлення центроїдів інвертованого файлового індексу, що використовується для наближеного пошуку найближчих сусідів у високовимірних векторних просторах. Показано, що запропоноване стохастичне правило оновлення безпосередньо мінімізує похибку реконструкції індексу та забезпечує поступову адаптацію структури розбиття до змін у розподілі векторів без потреби у повній реконструкції індексу, що є перевагою для масивів даних масштабу мільярдів записів.

5.1 Задача квантизації кольору

Запропонований метод SQC (4.12)-(4.13) може бути застосованим безпосередньо до широкого класу прикладних задач апроксимації дискретних розподілів. У цьому підрозділі ми розглядаємо конкретний випадок застосування

запропонованого методу стохастичної квантизації до задачі кольорової квантизації зображень – одного з підходів до стиснення зображень із втратами, що зберігає оригінальну роздільну здатність при суттєвому зменшенні обсягу інформації [94].

Цифрове представлення зображень у сучасних відеодисплеях базується на адитивному змішуванні кольорів, де інтенсивність трьох основних кольорів (червоного, зеленого та синього) модулюється незалежно у кожному пікселі [94]. У цій системі кожен піксель кодується трійкою беззнакових цілих чисел, де кожен елемент відповідає інтенсивності одного з основних кольорів. Зокрема, для 8-бітного дисплея кожен канал RGB набуває $2^8 = 256$ значень, що дає загалом $2^{8\times3} = 16777216$ можливих кольорів. Зі зростанням роздільної здатності сучасних зображень обсяг даних для їх зберігання та передавання суттєво зростає, що мотивує розробку ефективних алгоритмів стиснення, які забезпечують баланс між якістю зображення та обсягом даних.

Кольорова квантизація [94] є одним із підходів до стиснення зображень із втратами, що складається з двох послідовних етапів: побудови оптимальної колірної палітри домінантних кольорів зображення та відображення кожного пікселя на найближчий колір у цій палітрі. Цей підхід концептуалізує кожен піксель як точку у тривимірному просторі $\mathbb{R}^3$, де кожна вісь відповідає інтенсивності одного з трьох основних кольорів. Відповідно, пікселі зі схожими колірними відтінками розташовуються ближче один до одного в евклідовому просторі, що пов'язує задачу кольорової квантизації із задачею кластеризації. В роботі [61] було запропоновано розв'язання задачі кольорової квантизації із використанням алгоритму K-середніх [70] для побудови оптимальної колірної палітри, де центри кластерів відповідають оптимальним кольорам. Цей підхід був розвинутий у подальших дослідженнях [17,125].

Проте дослідження [92] виявило обмеження традиційних методів кластеризації при розв'язанні задач квантизації, зокрема їхню незадовільну масштабованість для великих наборів даних. З огляду на стрімке зростання роздільної здатності сучасних зображень, це обмеження масштабованості може суттєво вплинути на ефективність існуючих методів кольорової квантизації. У цьому контексті ми пропонуємо інтерпретацію задачі кольорової квантизації у термінах стохастичного програмування – як неопуклу стохастичну транспортну

задачу [64] – та застосовуємо алгоритм стохастичної квантизації [65,92] для побудови оптимальної колірної палітри.

Формально зведемо загальну задачу квантизації (4.3)-(4.4) до конкретного випадку кольорової квантизації. Нехай дискретний розподіл фіксованих елементів $\Xi = \{\xi_i,\, i = 1, \dots, N\}$ представляє множину пікселів оригінального зображення, де кожен елемент $\xi_i \in \mathbb{R}^3$ є трійкою інтенсивностей основних кольорів. Розподіл рухомих центрів $Y = \{y_k,\, k = 1, \dots, K\}$ відповідає оптимальній колірній палітрі, де кількість кольорів K задається як гіперпараметр задачі. Обидві множини $\{\xi_i\}$ та $\{y_k\}$ є підмножинами простору $\{(x, y, z) \in \mathbb{Z}^3 : 0 \le x, y, z \le 255\}$, що визначає комбіновану інтенсивність трьох основних кольорів для кожного пікселя зображення. Таким чином, задача кольорової квантизації є окремим випадком задачі (4.3) при $d = 3$, де розмірність простору визначається кількістю колірних каналів RGB.

Перед розв'язанням задачі оптимізації вхідний розподіл пікселів $\{\xi_i\}$ нормалізується до одиничного куба $[0, 1]^3 \subset \mathbb{R}^3$ шляхом ділення інтенсивності кожного каналу на максимальне значення 255. Як метрику відстані у цільовій функції (4.3) ми обираємо евклідову норму $\operatorname{dist}(\xi_i, y_k) = \|\xi_i - y_k\| = \sqrt{\sum_{l=1}^{3} |\xi_{il} - y_{kl}|^2}$, оскільки пікселі зі схожими кольорами мають близькі значення інтенсивностей основних кольорів і, відповідно, малу евклідову відстань між собою у просторі $\mathbb{R}^3$, а також подібність кольорів безпосередньо корелює з геометричною близькістю у просторі RGB [94].

Згідно з апроксимацією (4.11), задача кольорової квантизації зводиться до мінімізації неопуклої цільової функції (4.12)-(4.13) лише за позиціями центрів колірної палітри:

$$\min_{y=(y_1,\dots,y_K)\in[0,1]^{3K}} F(y) = \sum_{i=1}^{N} p_i \min_{1\le k\le K} \|\xi_i - y_k\|^r = \mathbb{E}_{i\sim p}\left[\min_{1\le k\le K} \|\xi_i - y_k\|^r\right] \tag{5.1}$$

де $p_i > 0$, $\sum_{i=1}^{N} p_i = 1$ – ймовірності пікселів (зазвичай рівномірні: $p_i = 1/I$), а допустима множина $Y = [0, 1]^3$ відповідає нормалізованому колірному кубу. Інтуїтивна інтерпретація задачі (5.1) полягає в наступному: для кожного пікселя оригінального зображення визначається найближчий колір в оптимальній палітрі, а цільова функція є зваженою сумою відстаней між пікселями та їхніми найближчими кольорами палітри. Мінімізація цієї функції забезпечує побудову

палітри, що найкраще апроксимує колірний розподіл оригінального зображення у сенсі метрики Васерштейна порядку r.

Для розв'язання задачі (5.1) ми застосовуємо алгоритм стохастичної квантизації SQC, описаний у попередньому підрозділі, з параметрами, специфічними для задачі кольорової квантизації: розмірність простору $d = 3$, допустима множина $Y = [0, 1]^3$ та параметр згладжування $r = 3$. Рекурентне правило оновлення позицій кольорів палітри має вигляд:

$$y_{k^*}^{(t+1)} = \Pi_{[0,1]^3}\left(y_{k^*}^{(t)} - \rho_t \cdot r\,\|\tilde{\xi}^{(t)} - y_{k^*}^{(t)}\|^{r-2}(y_{k^*}^{(t)} - \tilde{\xi}^{(t)})\right) \tag{5.2}$$

де $\tilde{\xi}^{(t)}$ – випадково обраний піксель оригінального зображення, $k^* \in \arg\min_{1\le k\le K} \|\tilde{\xi}^{(t)} - y_k^{(t)}\|$ – індекс найближчого кольору палітри, $\rho_t > 0$ – крок спуску, а $\Pi_{[0,1]^3}$ – оператор проекції на одиничний куб, що гарантує залишення кольорів палітри у допустимому діапазоні інтенсивностей.

На відміну від традиційних алгоритмів кластеризації, зокрема K-середніх [70], які на кожній ітерації оновлюють усі компоненти $y_k^{(t)}$ та використовують усі елементи $\{\xi_i\}$, запропонований підхід обробляє лише один елемент $y_{k^*}^{(t)}$ палітри та один випадковий піксель $\tilde{\xi}^{(t)}$ на кожній ітерації [92]. Просторова складність однієї ітерації становить $O(1)$, що лінійно залежить лише від розміру палітри та не залежить від кількості пікселів зображення N. Ця властивість усуває обмеження на об'єм оперативної пам'яті, необхідної для виконання оновлення, та забезпечує масштабованість для зображень високої роздільної здатності з мільйонами пікселів, де класичні методи потребують одночасного зберігання всіх N пікселів у пам'яті обчислювального пристрою з просторовою складністю $O(N)$. Крім того, у роботі [92] встановлено умови локальної збіжності алгоритму (5.2) за належних початкових умов, що зазвичай не гарантується для традиційних алгоритмів кластеризації.

З огляду на неопуклу цільову функцію в задачі (5.1), якість початкового наближення $y^{(0)} = \{y_k^{(0)},\, k = 1, \dots, K\}$ є визначальною для збіжності алгоритму до задовільного локального мінімуму. Хоча рівномірна випадкова вибірка із множини пікселів Ξ є найшвидшим підходом до ініціалізації, вона не забезпечує гарантій щодо якості початкового наближення та може призвести до збіжності алгоритму до суттєво субоптимального розв'язку.

Для побудови початкового наближення колірної палітри ми застосовуємо стратегію ініціалізації K-середніх++ [5], описану формулою (4.20), яка адаптується до контексту кольорової квантизації наступним чином:

1. вибрати перший колір палітри $y_1^{(0)}$ рівномірно випадково з множини пікселів Ξ;
2. для $s = 2, \dots, K$ обрати наступний колір $y_s^{(0)}$ з ймовірністю, пропорційно квадрату відстані від пікселя до найближчого з уже обраних кольорів:

$$P(\xi_j = y_s^{(0)}) = \frac{\min_{1 \le l < s} \|\xi_j - y_l^{(0)}\|^2}{\sum_{i=1}^{N} \min_{1 \le l < s} \|\xi_i - y_l^{(0)}\|^2} \tag{5.3}$$

3. повторювати крок 2 до отримання палітри $y^{(0)} = \{y_1^{(0)}, \dots, y_K^{(0)}\}$ з K кольорів.

Рівняння (5.3) показує, що пікселі, колір яких суттєво відрізняється від уже обраних кольорів палітри, мають вищу ймовірність бути включеними до початкової палітри. Це сприяє рівномірному покриттю колірного простору зображення та забезпечує $O(\log K)$-конкурентну апроксимацію оптимальної похибки квантизації [5]. Ефективність цієї стратегії ініціалізації для задачі стохастичної квантизації було емпірично підтверджено у роботі [92], де також проведено порівняння швидкості збіжності модифікованих версій алгоритму (5.2) з адаптивними кроками спуску.

Для оцінювання ефективності запропонованого методу стохастичної колірної квантизації ми провели серію порівняльних експериментів між алгоритмами SQC та K-середніх [70] на підмножині зображень із набору даних ImageNet [108], що зберігаються у форматі JPEG. З метою забезпечення єдиної масштабної бази для порівняння обчислювальних витрат всі зображення набору даних попередньо приведено до уніфікованої роздільної здатності 128×128 пікселів шляхом білінійної інтерполяції. Алгоритми протестовані для трьох розмірів оптимальної колірної палітри $K \in \{4, 8, 32\}$ у поєднанні з двома стратегіями побудови початкового наближення центрів: рівномірною випадковою вибіркою з множини пікселів Ξ та імовірнісною стратегією K-середніх++ [5], описаною формулою (5.3). Алгоритми реалізовано мовою Python із використанням бібліотеки NumPy [41] для ефективних тензорних операцій на CPU та бібліотеки Pillow [79] для завантаження, обробки та збереження зображень. Усі експерименти виконувались на віртуальній машині з двоядерним процесором та 8ГБ оперативної пам'яті. Для відтворюваності результатів зафіксовано генератор випадкових чисел

(seed $= 42$) та використано однакові значення гіперпараметрів ($\rho = 0.001$, $r = 3$) для всіх експериментів. Вихідний код та результати експериментів є загальнодоступними у репозиторії GitHub [58].

Оскільки запропонований алгоритм SQC оновлює позицію лише одного центру палітри $y_{k^*}^{(t+1)}$ на кожній ітерації згідно з правилом (5.2), тоді як алгоритм K-середніх [70] одночасно перераховує всі K центрів за повним проходом по множині пікселів, безпосереднє порівняння за кількістю ітерацій не відображає фактичних обчислювальних витрат методів. Натомість, інваріантною мірою обчислювальної складності, спільною для обох алгоритмів, є кількість обчислень функції відстані $\mathrm{dist}(\xi_i, y_k) = \|\xi_i - y_k\|$, оскільки саме ця операція домінує у часовій складності як процедури пошуку найближчого центру в SQC (5.2), так і кроку приписування пікселів до кластерів в алгоритмі K-середніх. У контексті задачі (6.1) кількість обчислень відстаней безпосередньо співвідноситься з обчислювальним часом, необхідним для побудови оптимальної колірної палітри. Як основну метрику якості квантизації для кількісного оцінювання подібності між оригінальним та стиснутим зображеннями ми використали метрику MSE. Результати порівняння алгоритмів SQC та K-середніх за кількістю обчислень функції відстані до моменту збіжності та за досягнутим значенням MSE наведено на рисунку 5.1.

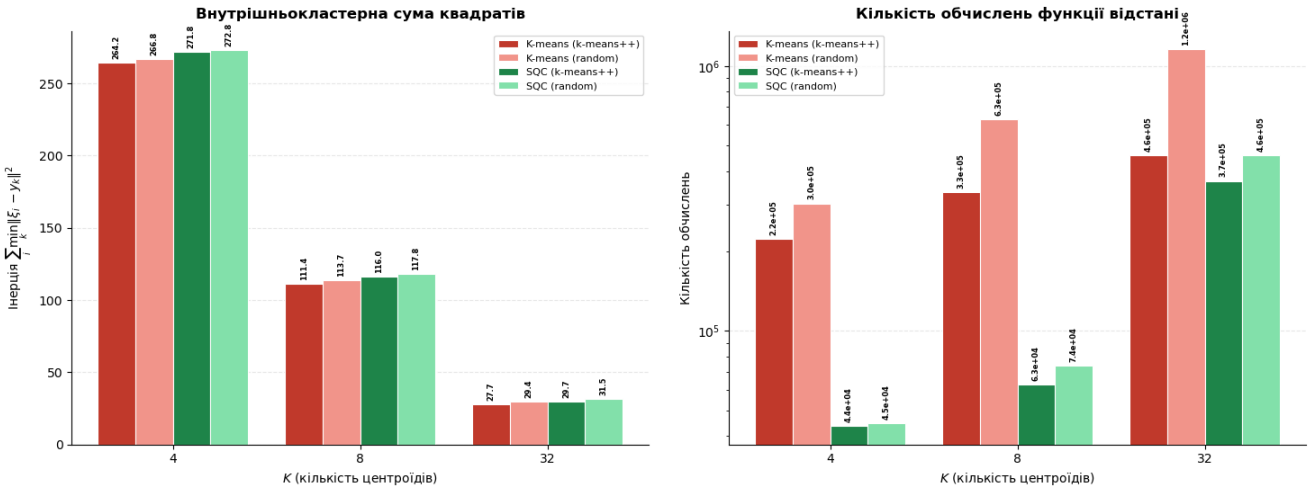

Рис. 5.1: Порівняння алгоритмів SQC та K-середніх [70] на підмножині набору даних ImageNet [108] (роздільна здатність 128×128 пікселів) за кількістю обчислень функції відстані та значенням MSE для розмірів палітри $K \in \{4, 8, 32\}$ та різних стратегій ініціалізації центрів.

Аналіз результатів, усереднених за всіма досліджуваними конфігураціями розміру палітри $K \in \{4, 8, 32\}$ та стратегіями ініціалізації центрів, демонструє, що запропонований метод SQC скорочує кількість обчислень функції відстані приблизно на 70% порівняно з алгоритмом K-середніх. При цьому значення MSE для обох алгоритмів є практично еквівалентними при фіксованому K, що свідчить про збереження якості квантизації за умови суттєвого зниження обчислювальних витрат. Очікувана монотонна залежність якості квантизації від розміру палітри зберігається для обох методів.

Для візуальної демонстрації ефективності алгоритму ми застосували його до зображення *ILSVRC2012_val_00023267* із набору даних ImageNet. Початкове зображення (роздільна здатність 1200×1206, розмір файлу 790.0КБ) було стиснуто до 207.4КБ з оптимальною палітрою із 8 кольорів. Візуальне порівняння оригінального та стиснутого зображень на рисунку 5.2 демонструє, що навіть при суттєвому зменшенні кількості кольорів основні структурні та тональні характеристики зображення зберігаються, а артефакти квантизації є мінімальними завдяки оптимальному розташуванню кольорів палітри у тривимірному колірному просторі. Підсумовуючи, результати чисельних експериментів дозволяють зробити три ключові висновки щодо ефективності стохастичної колірної квантизації:

1. запропонований метод забезпечує масштабоване розв'язання задачі кольорової квантизації без обмежень на об'єм пам'яті обчислювального пристрою та скорочує кількість обчислень функції відстані приблизно на 70% порівняно з алгоритмом K-середніх [70] без втрати якості квантизації (значення MSE є практично еквівалентними при фіксованому розмірі палітри K), що є перевагою порівняно з K-середніх, який потребує одночасного зберігання всіх N пікселів у пам'яті для виконання кожної ітерації оновлення центрів палітри;
2. стохастичні властивості алгоритму (5.2) дають можливість використання методів паралелізації для одночасного оновлення положень кількох кольорів палітри $\{y_k,\, k = 1, \dots, K\}$, що може забезпечити суттєве прискорення обчислень та є перспективним напрямком подальших досліджень;

3. швидкість збіжності алгоритму може бути додатково покращена шляхом заміни класичного стохастичного градієнтного спуску [32] адаптивними методами оновлення, такими як SQC-Adam, що автоматично підлаштовують крок оновлення для кожного кольору палітри окремо на основі накопиченої градієнтної інформації, як досліджено у роботі [92].

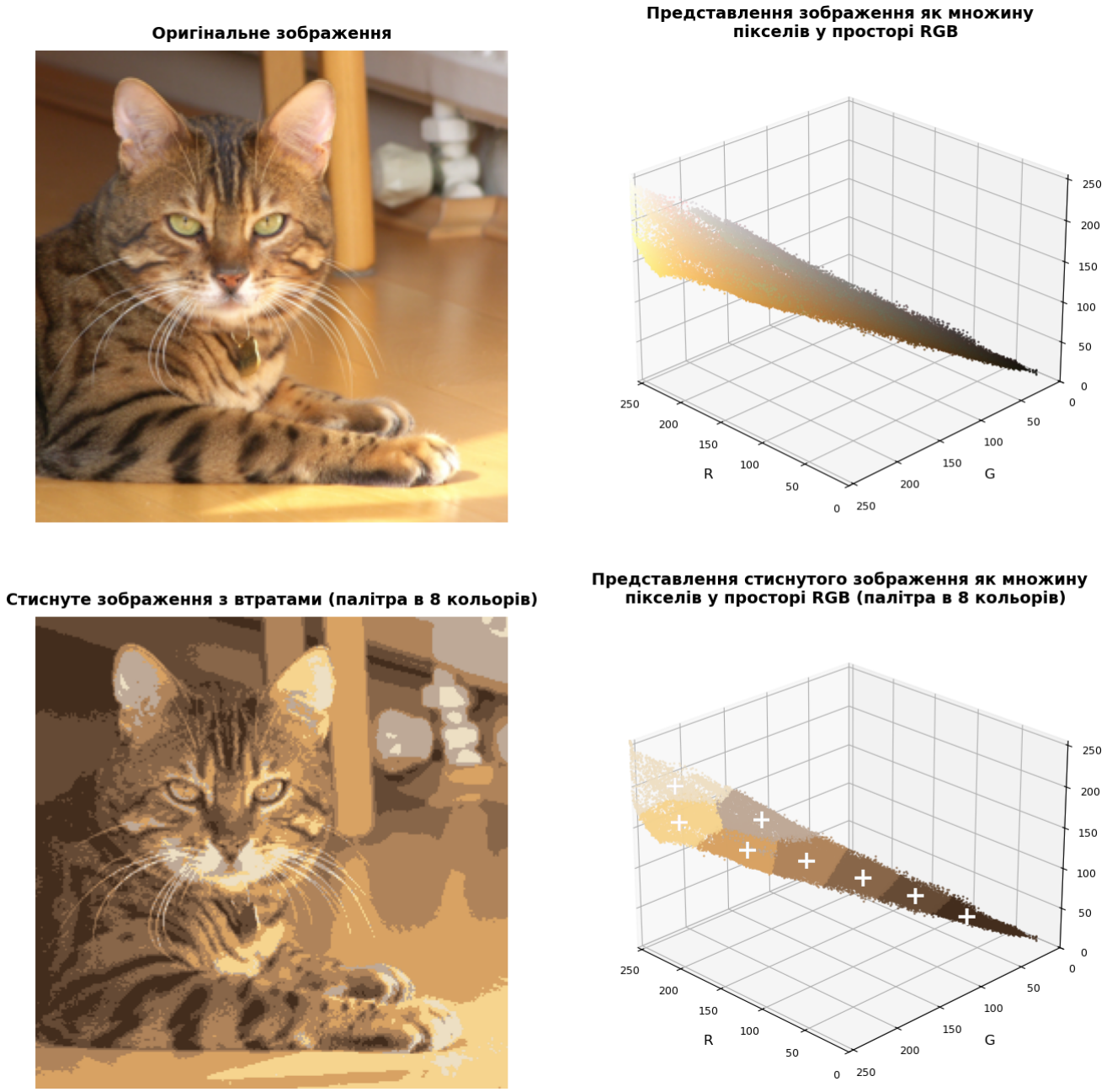

Рис. 5.2: Порівняння оригінального зображення до стиснутого зображення, отриманого запропонованим методом. Отримане зображення має меншу кількість кольорів в палітрі за початкове (8 кольорів), при тому зберігаючи основні візуальні характеристики.

Таким чином, запропонований метод не лише усуває обмеження існуючих методів кольорової квантизації, пов'язані з масштабованістю, але й відкриває нові можливості для оптимізації алгоритмів стиснення зображень у середовищах із обмеженими обчислювальними ресурсами.

5.2 Кодування даних за допомогою архітектури Triplet Network

Запропонований підхід класифікації долає обмеження методів кластеризації у високовимірних просторах шляхом інтеграції технік контрастивного навчання [42,54] з алгоритмом стохастичної квантизації, описаним у попередньому підрозділі. Хоча стохастичний алгоритм SQC із правилом оновлення (4.16) ефективно долає обмеження масштабованості для великих наборів даних за рахунок зменшення вимог до оперативної пам'яті з $O(Nd)$ до $O(d)$, він залишається вразливим до явища, відомого як «прокляття розмірності», – феномену, що є характерним для всіх методів кластеризації, заснованих на мінімізації відстані (зокрема, K-середніх та K-середніх++). В дослідженні [60] показано це явище та продемонстровано, що концепція відстані втрачає свою дискримінантну здатність у високовимірних просторах. Зокрема, розрізнення між найближчою та найвіддаленішою точками стає все менш значущим із зростанням розмірності d:

$$\lim_{d\to\infty} \frac{\max \operatorname{dist}(\xi_i, y_k) - \min \operatorname{dist}(\xi_i, y_k)}{\min \operatorname{dist}(\xi_i, y_k)} = 0 \tag{5.4}$$

Ця властивість є важливою для алгоритму SQC, оскільки операція пошуку найближчого центру $k^* \in S(\tilde{\xi}, y^t) = \arg\min_{1\le k\le K} \|\tilde{\xi} - y_k^t\|$, що виконується на кожній ітерації, стає неінформативною, коли відстані до всіх центрів стають практично однаковими. Як наслідок, стохастичні оцінки субградієнта (4.16) набувають шумового характеру, а збіжність алгоритму суттєво уповільнюється або порушується.

У цьому дослідженні розглядаються високовимірні дані у формі частково розмічених зображень рукописних цифр із набору даних MNIST [67]. Зазначимо, проте, що запропонований підхід не обмежується даними зображень та може бути застосований до інших типів високовимірних даних, таких як текстові документи

[101,127]. Хоча існують ефективні алгоритми зменшення розмірності, зокрема метод PCA [1,24], вони переважно призначені для відображення даних між двома неперервними просторами. Натомість, поставлена задача потребує алгоритму, що навчається ознакам подібності з дискретних наборів даних та проектує їх у метричний простір, де подібні елементи групуються у кластери, – процесу, відомого як навчання подібності.

Нещодавні дослідження [78,122] застосували архітектуру Triplet Network для навчання ознак з високовимірних дискретних даних зображень та їхнього кодування у низьковимірні представлення в латентному просторі. Автори запропонували підхід навчання, в якому трійкова мережа навчається на розміченій підмножині даних для кодування їх у латентні представлення в $\mathbb{R}^d$, а потім використовується для проектування решти нерозмічених даних у той самий латентний простір. Такий підхід суттєво зменшує витрати часу та праці на розмітку даних без суттєвого погіршення точності класифікації.

Трійкова мережа, запропонована у [42], є модифікацією загальної схеми контрастивного навчання [54]. Ключова ідея полягає в навчанні моделі з використанням трійок елементів:

1. випадковий елемент ξ_i: випадково обраний елемент з множини Ξ;
2. позитивний елемент ξ_i^+: елемент із міткою, що збігається з міткою якірного елемента ξ_i;
3. негативний елемент ξ_i^-: елемент із міткою, що відрізняється від мітки якірного елемента ξ_i.

На відміну від класичного контрастивного навчання, що порівнює лише пари позитивних та негативних елементів, трійкова мережа навчається мінімізувати відстань між випадковим та позитивним елементами, одночасно максимізуючи відстань між якірним та негативним елементами. Це досягається шляхом мінімізації цільової функції трійкових втрат (triplet loss):

$$\mathcal{L}_{triplet}(\theta) = \max\left(0,\ \mathrm{dist}(f_\theta(\xi_i), f_\theta(\xi_i^+)) - \mathrm{dist}(f_\theta(\xi_i), f_\theta(\xi_i^-)) + \alpha\right) \tag{5.5}$$

де $f_\theta : \Xi \to \mathbb{R}^d$ є параметризованим оператором відображення дискретних елементів Ξ у латентні представлення (у нашому випадку – трійкова мережа з вагами θ), $\mathrm{dist} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ є метрикою відстані між елементами, а $\alpha > 0$ є гіпер-параметром

дистанції (margin), що задає мінімальну відстань розділення між позитивними та негативними парами. За аналогією з метрикою відстані у цільовій функції стохастичної квантизації (4.12)-(4.13), для $\mathrm{dist}(\xi_i, \xi_j)$ у (6.5) використано евклідову норму l_2. Концептуальна діаграма цільової функції зображена на рисунку 5.3.

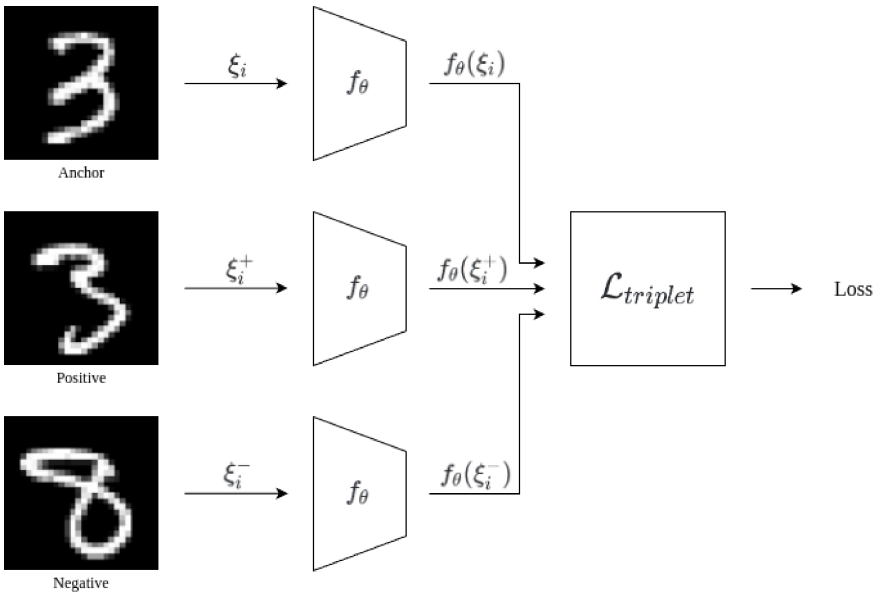

Рис. 5.3: Структура Triplet Network для набору даних MNIST [67].

У цьому дослідженні як параметризований оператор f_θ використано архітектуру згорткової DNN, запропоновану в [66]. АЗРП забезпечує ефективне обчислення градієнта цільової функції (5.5) за параметрами мережі θ шляхом послідовного застосування ланцюгового правила диференціювання від вихідного шару до вхідного [9,63,66]. На кожній ітерації навчання формується міні-пакет трійок $\{(\xi_i, \xi_i^+, \xi_i^-)\}_{i=1}^m$, для кожного елемента якого виконується прямий пропуск – послідовне обчислення активацій через шари $l = 1, \ldots, L$ мережі f_θ:

$$z^{[l]} = W^{[l]} a^{[l-1]} + b^{[l]}, \qquad a^{[l]} = \sigma_l\left(z^{[l]}\right) \tag{5.6}$$

де $a^{[0]} = \xi$ є вхідним зображенням, $W^{[l]}$ та $b^{[l]}$ – матриця ваг та вектор зміщень l-го шару, σ_l – функція активації. У прихованих шарах застосовується недиференційована в нулі функція ReLU: $\sigma_l(z) = \max(0, z)$; вихідний шар повертає латентне представлення без додаткового перетворення, тобто $f_\theta(\xi) = a^{[L]} \in \mathbb{R}^d$. Після прямого поширення для всіх трьох елементів трійки визначаються евклідові

відстані між латентними представленнями $d^+ = \|f_\theta(\xi_i) - f_\theta(\xi_i^+)\|$ та $d^- = \|f_\theta(\xi_i) - f_\theta(\xi_i^-)\|$, а також значення цільової функції трійкових втрат (5.5).

На етапі АЗРП градієнт $\nabla_\theta \mathcal{L}_{triplet}(\theta)$ обчислюється від вихідного шару до вхідного. Для активних трійок (де $d^+ - d^- + \alpha > 0$) сигнал помилки вихідного шару – градієнт цільової функції за латентним представленням якірного елемента – набуває вигляду:

$$\delta^{[L]} = \frac{\partial \mathcal{L}_{triplet}}{\partial f_\theta(\xi_i)} = \frac{f_\theta(\xi_i) - f_\theta(\xi_i^+)}{d^+} - \frac{f_\theta(\xi_i) - f_\theta(\xi_i^-)}{d^-} \tag{5.7}$$

Для неактивних трійок ($d^+ - d^- + \alpha \leq 0$) сигнал помилки $\delta^{[L]} = 0$, і внесок цієї трійки в оновлення параметрів мережі є нульовим. Аналогічні сигнали помилки обчислюються для позитивного та негативного елементів трійки з відповідними протилежними знаками. Далі $\delta^{[L]}$ послідовно поширюється назад через приховані шари:

$$\delta^{[l]} = \left(W^{[l+1]\top} \delta^{[l+1]}\right) \odot \sigma_l'\left(z^{[l]}\right), \quad l = L-1, \ldots, 1 \tag{5.8}$$

де $\odot$ позначає поелементне множення, а σ_l' – субградієнт функції активації шару l. Для функції ReLU $\sigma'(z) = \mathbf{1}[z > 0]$ в точках $z \neq 0$; у точці $z = 0$ використовується субградієнтне узагальнення АЗРП на негладкий неопуклий випадок, теоретично обґрунтоване у [86]. На основі сигналів помилки $\{\delta^{[l]}\}$ та активацій прямого проходу $\{a^{[l]}\}$ обчислюються градієнти цільової функції за параметрами кожного шару:

$$\frac{\partial \mathcal{L}_{triplet}}{\partial W^{[l]}} = \delta^{[l]} \left(a^{[l-1]}\right)^\top, \qquad \frac{\partial \mathcal{L}_{triplet}}{\partial b^{[l]}} = \delta^{[l]}, \quad l = 1, \ldots, L \tag{5.9}$$

Обчислені градієнти (5.9) використовуються для оновлення параметрів мережі θ методом адаптивної оцінки моментів Adam (2.58). Для стохастичної оцінки градієнта $g^k = \nabla_\theta \mathcal{L}_{triplet}(\theta^k)$, обчисленої за міні-пакетом трійок розміру m, правило оновлення набуває вигляду:

$$\theta^{k+1} = \theta^k - \frac{\rho}{\sqrt{\bar{v}^k + \varepsilon}}\, \bar{m}^k \tag{5.10}$$

де $\bar{m}^k = m^k/(1-\beta_1^k) \quad \bar{v}^k = v^k/(1-\beta_2^k)$ є незміщеними оцінками першого та другого моментів градієнта відповідно [56]. Ця стратегія навчання є особливо придатною для задачі мінімізації трійкових втрат із стратегією напівжорсткого вибору (5.11), де значна частина трійок може бути неактивними на певній ітерації і давати нульовий вклад у градієнт, що призводить до розрідженості стохастичних оцінок

g^k та вимагає адаптивного підлаштування ефективного кроку навчання для кожного параметра окремо згідно з накопиченим станом другого моменту $\bar{v}^k$.

Щодо стратегій вибору трійок $(\xi_i, \xi_i^+, \xi_i^-)$, визначальним є обрання підходу, що генерує найбільш інформативні градієнти для цільової функції (6.5). Робота [130] аналізує різні стратегії ітеративного вибору трійок, що формують трійки в межах кожного міні-пакету навчальної множини на кожній ітерації. У цьому дослідженні застосовано стратегію напівжорсткого вибору трійок (semi-hard triplet mining), що обирає негативний елемент, який знаходиться далі від якірного, ніж позитивний елемент, проте в межах дистанції α:

$$\xi_i^- = \arg \min_{\substack{\xi : C(\xi) \neq C(\xi_i) \\ \mathrm{dist}(f_\theta(\xi_i), f_\theta(\xi)) > \mathrm{dist}(f_\theta(\xi_i), f_\theta(\xi_i^+))}} \mathrm{dist}(f_\theta(\xi_i), f_\theta(\xi)) \tag{5.11}$$

де $C(\xi)$ позначає мітку класу елемента ξ. Інтуїтивна інтерпретація стратегії (5.11) полягає в наступному: вибираються негативні елементи, що є достатньо складними для розрізнення (знаходяться ближче до якірного, ніж передбачає дистанцію α), але не настільки складними, щоб спричинити нестабільність навчання, що забезпечує збалансоване зростання якості латентних представлень.

Застосовуючи ідеї з [42,78,122], спроектовані латентні представлення Triplet Network використовуються для навчання алгоритму стохастичної квантизації з правилом оновлення (4.16) та його адаптивними модифікаціями. Цей комбінований підхід дозволяє розв'язувати задачі керованої або класифікації високовимірних даних. Процес навчання «без учителя» за допомогою комбінованого алгоритму, за умови наявності розміченої підмножини $\Xi' \subset \Xi$ та решти нерозмічених даних $\bar{\Xi} = \Xi \setminus \Xi'$, може бути узагальнений наступним чином:

1. навчання Triplet Network f_θ на розмічених даних Ξ' та побудова їхніх спроектованих латентних представлень у $\mathbb{R}^d$;
2. використання навченої Triplet Network f_θ для проектування решти нерозмічених даних $\bar{\Xi}$ у той самий латентний простір представлень $\mathbb{R}^d$;
3. навчання алгоритму стохастичної квантизації SQC з правилом оновлення (4.16) на об'єднанні розмічених та нерозмічених латентних представлень.

Таким чином, трійкова мережа виконує роль проміжного відображення, що перетворює високовимірні дискретні дані в низьковимірні неперервні представлення, у яких евклідова відстань зберігає семантичну подібність

елементів, а алгоритм SQC здійснює оптимальну квантизацію отриманих латентних представлень, уникаючи проблеми (5.4) втрати дискримінантної здатності відстані.

Для реалізації та навчання Triplet Network використано фреймворк PyTorch 2.0 [4], призначений для масштабованих паралельних обчислень на спеціалізований обчислювальних пристроях. Алгоритм стохастичної квантизації було реалізовано з використанням високорівневого програмного інтерфейсу бібліотеки Scikit-learn [95], що забезпечує сумісність з іншими компонентами пакету (зокрема, навчанням та оцінюванням моделей), із застосуванням бібліотеки NumPy [41] для ефективних тензорних обчислень на центральному процесорі. Усі рисунки, представлені у цьому дослідженні, побудовано за допомогою бібліотеки Matplotlib [45]. Вихідний код та результати експериментів є у відкритому доступі у репозиторії GitHub [58].

Для проведення експериментів використано оригінальний набір даних MNIST рукописних цифр [67], що містить 60000 напівтонових зображень рукописних цифр із роздільною здатністю 28×28 пікселів, кожне з яких пов'язане з міткою класу від 0 до 9. Додатково набір даних включає відповідну тестову множину з 10000 зображень та їхніми мітками. Зазначимо, що жодних методів доповнення (augmentation) або попередньої обробки не було застосовано ні до навчальних, ні до тестових даних, що дозволяє оцінити ефективність запропонованого підходу безпосередньо на вихідних даних. Приклади вхідних даних набору MNIST зображені на рисунку 5.4.

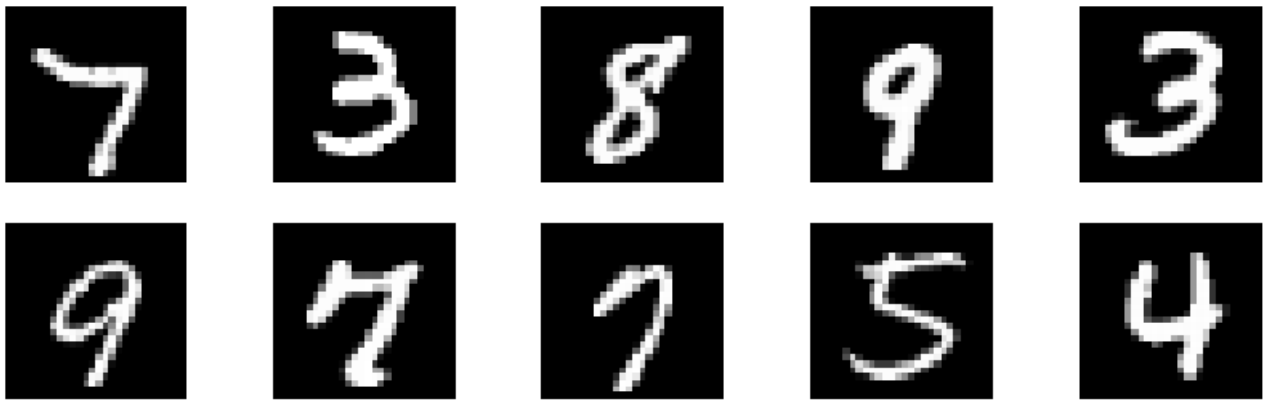

Рис. 5.4: Репрезентативні зразки з набору даних MNIST [67] із відповідними мітками класів.

Задачу класифікації зображень сформульовано як задачу навчання «без учителя» з навчанням моделей на різних частках розмічених навчальних даних (25%, 50%, 75% та 100%). Навчальний набір даних розбивався шляхом рівномірної вибірки на розмічену та нерозмічену частини відповідно до зазначених відсотків. Для Triplet Network використано архітектуру згорткової DNN, що складається з двох згорткових шарів із фільтрами розміру 3×3 (розмірності карт ознак 32 та 64 відповідно, з параметрами stride = 1 та padding = 1, за якими слідують шари максимального об'єднання (max-pooling) розміру 2×2, та двох повнозв'язних шарів. У всіх шарах мережі використано недиференційовні функції активації ReLU, що вносять нелінійність у кожен шар. Разом із негладкою цільовою функцією трійкових втрат (5.5) це робить задачу навчання суттєво негладкою та неопуклою [86].

Для кожної частки розмічених даних навчено окрему модель Triplet Network з наступними гіпер-параметрами: 50 епох навчання, розмір міні-пакету $m = 1000$, крок навчання $\rho = 10^{-3}$, коефіцієнт l_2-регуляризації $\lambda = 10^{-5}$. Для цільової функції трійкових втрат (5.5) та стратегії вибору трійок (5.11) встановлено гіпер-параметр дистанції $\alpha = 1.0$. Для забезпечення змістовного виявлення ознак із одночасною можливістю візуалізації обрано латентний простір $\mathbb{R}^3$ розмірності $d = 3$. На рисунку 5.5 представлено латентні представлення зображень навчального набору даних (ліворуч) та тестового набору даних (праворуч), проектовані трійковою мережею, де кожен елемент має колір відповідно до його мітки (0-9). Кластеризація елементів із однаковою міткою свідчить про те, що трійкова мережа успішно виявила релевантні ознаки під час навчання, а латентний простір зберігає семантичну структуру вхідних даних.

Трійкову мережу використано для проектування латентних представлень у простір $\mathbb{R}^3$ як розмічених, так і нерозмічених навчальних даних. Отримані представлення потім використано для навчання алгоритму стохастичної квантизації SQC як моделі некерованого навчання. Для кожного набору латентних

представлень, що відповідає різним часткам розмічених даних, навчено алгоритм SQC та його адаптивні варіанти на основі методів (2.48), (2.52), (2.58).

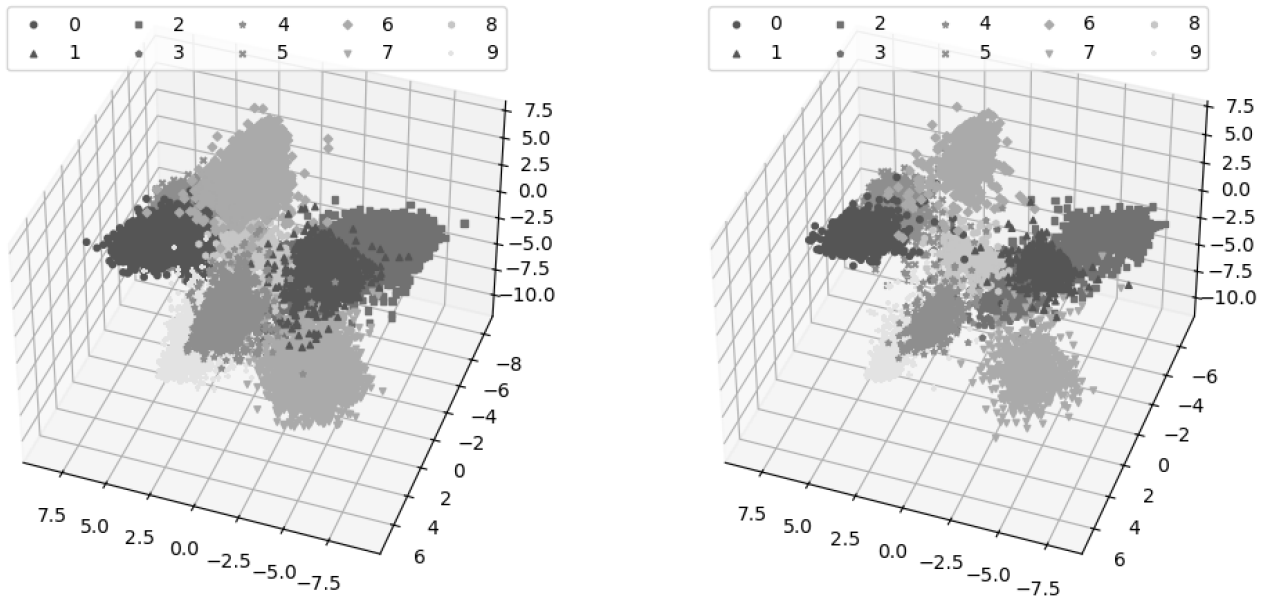

Рис. 5.5: Латентні представлення зображень навчального набору даних (ліворуч) та тестового набору даних (праворуч), проектовані трійковою мережею, з кольоровим кодуванням за мітками класів (0-9). Кластеризація елементів із однаковою міткою свідчить про успішне виявлення релевантних ознак під час навчання.

Стратегія ініціалізації позицій центрів використовувала рівномірну вибірку з розміченої частки даних для кожної мітки класу. Для забезпечення збіжності встановлено параметр згладжування $r = 3$ та застосовано різні кроки навчання для кожного варіанту: $\rho = 0.001$ для SQC, SQC-Momentum, SQC-NAG, SQC-RMSProp; $\rho = 0.1$ для SQC-AdaGrad; та $\rho = 0.01$ для SQC-Adam. З цими гіпер-параметрами всі варіанти стохастичної квантизації збіглися до глобальних оптимумів.

На рисунку 5.6 представлено оптимальні позиції квантів для кожного варіанту стохастичної квантизації (позначених назвою варіанту) у латентному просторі відносно розмічених елементів (0-9) тестового набору даних. Візуальний аналіз підтверджує, що всі варіанти алгоритму SQC успішно визначають геометричне розташування центрів квантизації, які відповідають кластерній структурі латентних представлень. На рисунку 5.7 подано порівняльний аналіз

швидкості збіжності варіантів стохастичної квантизації шляхом побудови графіку значення цільової функції (4.13) залежно від номера ітерації t.

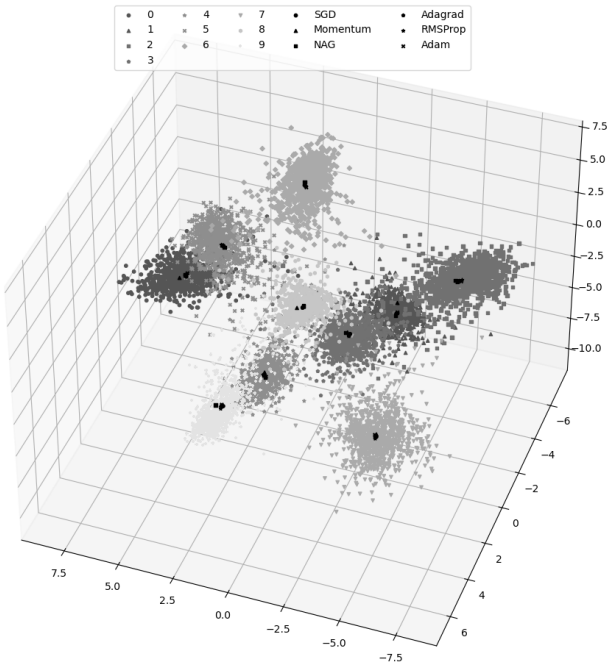

Рис. 5.6: Оптимальні позиції квантів для кожного варіанту стохастичної квантизації (позначених назвою варіанту) у латентному просторі відносно розмічених елементів (0-9) тестового набору даних.

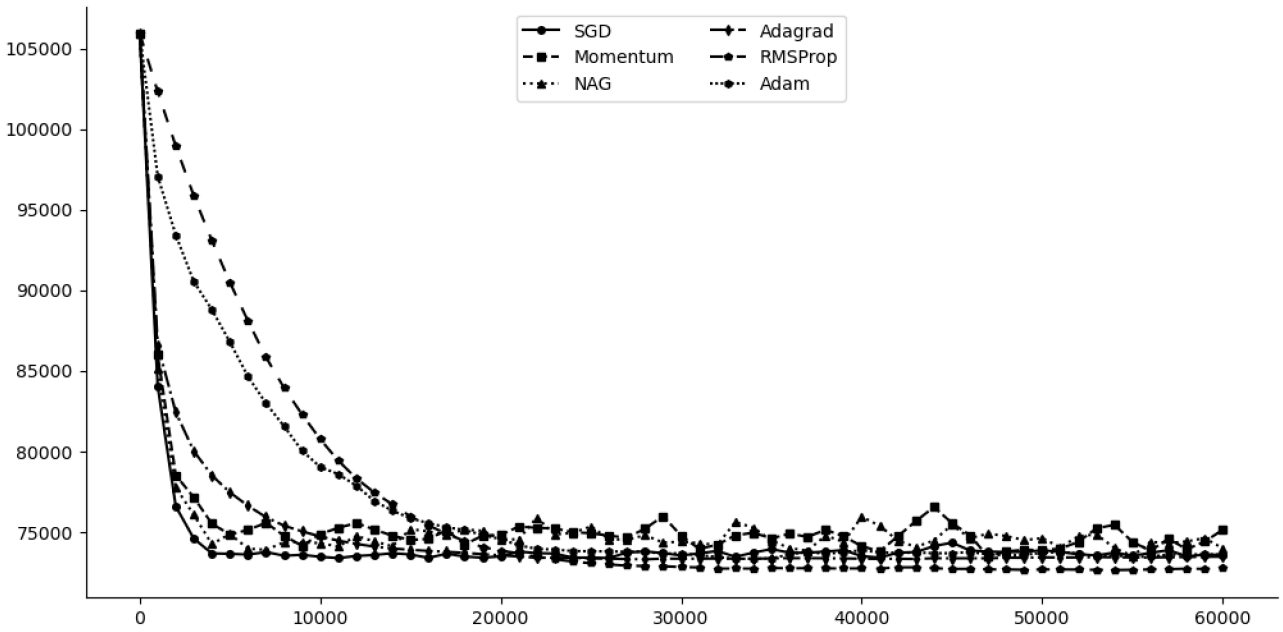

Рис. 5.7: Порівняння швидкості збіжності варіантів стохастичної квантизації на латентних представленнях навчальних даних із 100% часткою розмічених даних.

Аналіз графіків збіжності дозволяє зробити кілька суттєвих висновків щодо динаміки оптимізації. По-перше, адаптивні варіанти SQC-RMSProp та SQC-Adam демонструють найшвидше зменшення значення цільової функції на початкових ітераціях, що підтверджує ефективність покоординатної адаптації кроку оновлення, описаної в попередньому підрозділі. По-друге, варіант SQC-AdaGrad із більшим початковим кроком $\rho = 0.1$ забезпечує порівнянну швидкість збіжності, проте демонструє характерне для цього методу монотонне зменшення ефективного кроку внаслідок накопичення квадратів градієнтів, що уповільнює збіжність на фінальних ітераціях. По-третє, базовий алгоритм SQC та його варіанти з імпульсом (SQC-Momentum, SQC-NAG) демонструють стабільну, проте повільнішу збіжність, що зумовлено застосуванням єдиного скалярного кроку до всіх компонент вектора оновлення. Отримані результати узгоджуються з теоретичними оцінками швидкості збіжності методу SQC [76] та підтверджують практичну ефективність адаптивних модифікацій для задач стохастичної квантизації на реальних даних [58].

Точність кластеризації навчених моделей, що поєднують трійкову мережу та стохастичну квантизацію, оцінено за допомогою індексу Ранда [102], обчисленого відносно «істинних» міток класів (0-9) тестового набору даних MNIST [67]. Нехай $X = \{X_1, \dots, X_K\}$ є розбиттям тестової вибірки $\Xi^{\text{test}} = \{\xi_i\}_{i=1}^{N^{\text{test}}}$ на K кластерів, отриманим запропонованим алгоритмом стохастичної квантизації шляхом приписування кожного елемента ξ_i до найближчого центру y_{k^*}, а $Y = \{Y_1, \dots, Y_K\}$ – еталонне розбиття тієї ж вибірки за «істинними» мітками класів MNIST. Для кожної невпорядкованої пари елементів (ξ_i, ξ_j), $i < j$, визначимо чотири взаємовиключні випадки: a – кількість пар, що належать одному кластеру у обох розбиттях X та Y; b – кількість пар, що належать різним кластерам у обох розбиттях; c – кількість пар, що належать одному кластеру у X, але різним кластерам у Y; d – кількість пар, що належать різним кластерам у X, але одному кластеру у Y. Індекс Ранда визначається як частка узгоджених рішень про парне групування серед усіх можливих пар [102]:

$$\mathrm{RI}(X, Y) = \frac{a + b}{a + b + c + d} = \frac{a + b}{\binom{N^{\text{test}}}{2}} \in [0, 1] \tag{5.12}$$

де значення $\mathrm{RI} = 1$ відповідає повному збігу розбиттів X та Y (з точністю до перестановки кластерних міток), а $\mathrm{RI} = 0$ – повній неузгодженості. Індекс (5.12) вимірює, з якою ймовірністю запропонований алгоритм кластеризації приймає те саме рішення щодо двох випадково обраних елементів (приписати їх до одного кластера або до різних), що й еталонне розбиття за «істинними» мітками класів MNIST. Результати експериментів продемонстрували, що запропонований підхід досягає значень індексу Ранда, порівнянних із найкращими відомими результатами для Triplet Network та Siamese Network, повідомленими у [42], навіть за умови обмеженої кількості розмічених даних. У таблиці 5.1 представлено вичерпне порівняння значень індексу Ранда для різних алгоритмів та часток розмічених навчальних даних.

Табл. 5.1: Порівняння точності кластеризації за індексом Ранда (6.12) [102] на наборі даних MNIST [67].

Алгоритм		Частка розмічених навчальних даних			
		25%	50%	75%	100%
Siamese Network	KNN	-	-	-	0.979
Triplet Network	KNN	-	-	-	***0.9954***
	SQC-SGD	0.781	***0.9774***	0.9783	0.9838
	SQC-Momentum	0.7806	0.9773	0.9782	0.9836
	SQC-NAG	0.78	0.9763	0.9772	0.9833
	SQC-AdaGrad	0.7895	0.9772	0.9779	0.9834
	SQC-RMSProp	***0.7934***	0.977	0.977	0.9829
	SQC-Adam	0.7915	0.977	***0.9786***	0.9829

При зменшенні частки розмічених даних до 50% запропонований підхід зберігає високу узгодженість із еталонним розбиттям ($\mathrm{RI} = 0.9774$ для базового SQC), що

свідчить про ефективність запропонованого гібридного методу навчання «без учителя».

Інтеграція трійкової мережі з алгоритмом стохастичної квантизації SQC ефективно долає «прокляття розмірності» (5.4) шляхом відображення високовимірних дискретних даних ($28 \times 28 = 784$ виміри) у низьковимірний латентний простір ($d = 3$), де евклідова відстань зберігає кластерну структуру та дискримінантну здатність. Адаптивні модифікації алгоритму SQC – зокрема SQC-Adam та SQC-RMSProp – забезпечують прискорену збіжність та підвищене значення індексу Ранда (5.12) порівняно з базовим стохастичним субградієнтним спуском, особливо в режимах обмеженої розмітки даних. Ці результати мотивують подальше застосування запропонованого комбінованого підходу до більш складних задач класифікації високовимірних даних, зокрема у сценаріях із обмеженими обчислювальними ресурсами, де незалежна від розміру вибірки N просторова складність $O(Kd)$ алгоритму SQC стає суттєвою перевагою [58,64].

5.3 Стохастичний інвертований файловий індекс

Запропонований метод SQC має прикладне застосування у задачі оновлення індексних структур для наближеного пошуку найближчих сусідів у високовимірних просторах. У цьому підрозділі ми показуємо, що алгоритм SQC із правилом оновлення (4.16) та його адаптивними модифікаціями може бути безпосередньо застосований як заміна алгоритму K-середніх для ітеративного оновлення центроїдів IVF, усуваючи потребу у повній перебудові індексу при зміні розподілу елементів множини вхідних даних.

Інвертований файловий індекс, запропонований у роботі [49], є розподільною структурою даних, що адресує обчислювальні складності пошуку за подібністю у масштабі шляхом розбиття простору на n_c непересічних областей та групуванні елементів у інвертовані списки відповідно до їхніх призначень до розділів. Формально, IVF індекс над множиною елементів $\mathcal{X} = \{x_i \in \mathbb{R}^d,\, i = 1, \dots, N\}$ будується за допомогою методу квантизації $q_1 : \mathbb{R}^d \to \mathcal{C}_1 \subset \mathbb{R}^d$, де розбиття Вороного $\mathcal{C}_1 = \{c_j \in \mathbb{R}^d,\, j = 1, \dots, n_c\}$ містить n_c центроїдів, визначених алгоритмом

кластеризації [48]. Найбільш поширений підхід застосовує кластеризацію K-середніх для визначення оптимальних позицій центроїдів:

$$\mathcal{C}_1 = \arg \min_{\{c_1,\dots,c_{n_c}\}} \sum_{i=1}^{N} \min_{j=1,\dots,n_c} \|x_i - c_j\|^2 \tag{5.13}$$

що визначає розбиття простору на n_c кластерів із відповідними інвертованими списками $\mathcal{I}_j = \{x_i \in \mathcal{X} : q_1(x_i) = c_j\}$, $j = 1, \dots, n_c$ [49]. Кількість центроїдів зазвичай обирається як $n_c \approx \sqrt{N}$ для збалансування гранулярності розбиття з обчислювальними витратами пошуку серед центроїдів [48].

Пошук у IVF індексі реалізує двоетапну стратегію фільтрації для даного запиту $q \in \mathbb{R}^d$: спочатку визначаються n_p найближчих центроїдів $\mathcal{L}_{\text{IVF}} = n_p\text{-}\arg\min_{c \in \mathcal{C}_1} \|q - c\|^2$, а потім виконується точний пошук серед елементів відповідних інвертованих списків [48]. Параметр n_p забезпечує явний контроль компромісу між якістю пошуку та обчислювальною вартістю: більші значення n_p підвищують повноту вибірки за рахунок сканування більшої кількості розділів, проте збільшують обчислювальне навантаження [48]. На практиці IVF індекси часто поєднуються з PQ для стиснення векторних представлень, утворюючи варіант IVFADC, де залишкові елементи $r = x - q_1(x)$ стискуються за допомогою методу квантизації добутку q_2 із попередньо обчисленими таблицями відстаней [49].

Застосування методу SQC у контексті IVF індексу мотивоване структурною відповідність між задачею побудови розбиттів Вороного (5.13) та задачею стохастичної квазіградієнтної кластеризації (4.12)-(4.13). Нехай множина елементів $\mathcal{X} = \{x_i,\, i = 1, \dots, N\}$ ідентифікується з множиною фіксованих елементів $\{\xi_i,\, i = 1, \dots, N\}$ за $\xi_i = x_i$, центроїди IVF індексу $\mathcal{C}_1 = \{c_1, \dots, c_{n_c}\}$ – з рухомими центрами $\{y_1, \dots, y_K\}$ при $K = n_c$, а ймовірності елементів покладаються рівномірними: $p_i = 1/N$. Тоді задача (5.13) є окремим випадком задачі (4.12)-(4.13) при $r = 2$ та евклідовій нормі $\text{dist}(\xi_i, y_k) = \|\xi_i - y_k\|$:

$$\min_{\{c_1,\dots,c_{n_c}\}} \sum_{i=1}^{N} \min_{j=1,\dots,n_c} \|x_i - c_j\|^2 = N \cdot \min_{y \in \mathbb{R}^{d \cdot n_c}} \frac{1}{N} \sum_{i=1}^{N} \min_{1 \le k \le n_c} \|x_i - y_k\|^2 = N \cdot F(y) \tag{5.14}$$

де $F(y) = \mathbb{E}_{i \sim p}[\min_{1 \le k \le n_c} \|x_i - y_k\|^2]$ є цільовою функцією задачі стохастичної квантизації. Таким чином, побудова IVF індексу є еквівалентною задачі стохастичної квазіградієнтної кластеризації дискретного рівномірного розподілу

на множині елементів, а алгоритм K-середніх є окремим випадком декомпозиційного алгоритму етапу 2 при $r = 2$. Ця еквівалентність дозволяє застосовувати усі теоретичні відомості стохастичної апроксимації задачі (4.13), на задачу оновлення IVF індексу.

Проте за умов динамічного середовища надходження даних, коли нові елементи додаються та наявні видаляються з індексу, якість IVF деградує внаслідок дисбалансу розділів та зростання похибки реконструкції. Дисбаланс розділів виникає, коли розподіл розмірів інвертованих списків суттєво відхиляється від початкового збалансованого стану, що призводить до змінної латентності запитів. Похибка реконструкції, визначена як середня відстань між елементами та їхніми призначеними центроїдами $\bar{\varepsilon} = \frac{1}{N}\sum_{i=1}^{N} \|x_i - q_1(x_i)\|^2$, зростає у міру того, як кластерна структура відхиляється від оптимального розбиття, визначеного під час побудови індексу [16,48]. На практиці дані проблеми овна перебудова індексу, що є ефективною для відновлення оптимального розбиття, але потребує багато обчислювальних ресурсів для масивів даних масштабу мільярдів записів.

Існуючі підходи до оновлення IVF індексу можна класифікувати за рівнем модифікації структури розбиттів. Найпростіший підхід зберігає фіксовані центроїди та обмежується додаванням нових елементів до відповідних інвертованих списків за правилом $c^* = \arg\min_{c \in \mathcal{C}_1} \|x_{\text{new}} - c\|^2$. Цей підхід є обчислювально ефективним, проте не адаптує структуру індексу до змін у розподілі елементів. Більш адаптивним є підхід оновлення центроїдів з перерахуванням розбиття як середнє всіх елементів, поточно призначених до нього:

$$c_j \leftarrow \frac{1}{|\mathcal{I}_j|} \sum_{x \in \mathcal{I}_j} x \tag{5.15}$$

що може бути обчислено ефективно за допомогою накопичувальних статистичних моментів. Проте підхід (6.15) мінімізує внутрішньокластерну дисперсію лише для фіксованого призначення елементів до розбиттів. Як наслідок, елементи, що були б більш оптимально призначені до іншого розбиття після зміщення центроїдів, залишаються у попередніх інвертованих списках, що обмежує зменшення похибки реконструкції.

Запропонований метод SQC усуває зазначені обмеження завдяки властивостям стохастичної апроксимації, що дозволяє одночасно оновлювати позиції центроїдів та неявно коригувати призначення окремих елементів до розбиття без повного перерахунку. Зведення, встановлене рівністю (5.14), дозволяє безпосередньо застосувати стохастичне правило оновлення (4.13) до центроїдів IVF індексу. При надходженні нового елемента x_{new} алгоритм виконує два одночасних кроки:

1. звичайне додавання елемента до відповідного інвертованого списку;
2. стохастичне оновлення розміщення найближчого центроїда у напрямку зменшення цільової функції (5.14).

Формально, рекурентне правило оновлення центроїдів IVF індексу при надходженні нового елемента x_{new} набуває вигляду:

$$c_{k^*}^{(t+1)} = c_{k^*}^{(t)} - \rho_t \cdot r \, \|x_{\text{new}} - c_{k^*}^{(t)}\|^{r-2} \left(c_{k^*}^{(t)} - x_{\text{new}}\right) \tag{5.16}$$

де $k^* = \arg\min_{1 \le j \le n_c} \|x_{\text{new}} - c_j^{(t)}\|$ є індексом найближчого центроїда, $\rho_t > 0$ є кроком спуску, а $r \ge 1$ є параметром згладжування. Інтерпретація правила (5.16) полягає в фізичному явищі, де кожен новий елемент, що надходить до індексу, «притягує» найближчий центроїд у своєму напрямку з силою, пропорційною ступеню відстані $\|x_{\text{new}} - c_{k^*}^{(t)}\|^{r-2}$. При $r = 2$ оновлення (5.16) набуває вигляду $c_{k^*}^{(t+1)} = c_{k^*}^{(t)} - 2\rho_t(c_{k^*}^{(t)} - x_{\text{new}})$, що є лінійним зміщенням центроїда у напрямку нового елемента. При $r = 3$, рекомендованому на основі результатів з дослідження методу узагальненого градієнтного спуску, субградієнт масштабується відстанню, що забезпечує більш виражене коригування центроїдів для елементів, розташованих далеко від поточних центрів, та поступове згладжування при наближенні розподілу до стаціонарного стану.

Повний алгоритм ітеративного оновлення IVF індексу на основі методу SQC формулюється наступним чином. Задано IVF індекс із центроїдами $\mathcal{C}_1^{(0)} = \{c_1^{(0)}, \dots, c_{n_c}^{(0)}\}$ та інвертованими списками $\{\mathcal{J}_1, \dots, \mathcal{J}_{n_c}\}$, побудований за вихідною множиною $\mathcal{X}$, та гіперпараметри: крок спуску $\rho > 0$, параметр згладжування $r = 3$, коефіцієнти статистичних моментів $\beta_1, \beta_2 \in (0, 1)$.

1. Ініціалізація: центроїди $\mathcal{C}_1^{(0)}$ з побудови індексу, статистичні моменти $m_j^0 = 0$, $v_j^0 = 0$ для всіх $j = 1, \dots, n_c$;

2. При надходженні нового вектора x_{new} на ітерації t:
   i. пошук найближчого центроїда: $k^* = \arg\min_{1 \le j \le n_c} \|x_{\text{new}} - c_j^{(t)}\|$;
   ii. додавання елемента до інвертованого списку: $\mathcal{J}_{k^*} \leftarrow \mathcal{J}_{k^*} \cup \{x_{\text{new}}\}$;
   iii. обчислення стохастичної оцінки субградієнта: $g_{k^*}^t = r\,\|x_{\text{new}} - c_{k^*}^{(t)}\|^{r-2}(c_{k^*}^{(t)} - x_{\text{new}})$;
   iv. оновлення оцінки першого моменту: $m_{k^*}^t = \beta_1\, m_{k^*}^{t-1} + (1-\beta_1)\, g_{k^*}^t$;
   v. оновлення оцінки другого моменту: $v_{k^*}^t = \beta_2\, v_{k^*}^{t-1} + (1-\beta_2)\, \|g_{k^*}^t\|^2$;
   vi. оновлення позиції центроїда: $c_{k^*}^{(t+1)} = c_{k^*}^{(t)} - \rho_t/(\sqrt{v_{k^*}^t + \varepsilon})\, m_{k^*}^t$
3. При видаленні елемента x_{del} з індексу: виконати аналогічне оновлення центроїда з протилежним знаком субградієнта.

Обчислювальна складність кожної операції вставки з одночасним оновленням центроїда становить $O(n_c d)$, де крок (i) потребує обчислення n_c відстаней у $\mathbb{R}^d$, а кроки (iii)-(vi) виконуються за час $O(d)$. Ця складність є ідентичною складності стандартної вставки у IVF індекс із фіксованими центроїдами, оскільки крок (i) присутній в обох підходах, а додаткові кроки оновлення центроїда мають нехтовно малу складність $O(d)$ порівняно з $O(n_c d)$.

На відміну від рівняння (5.15), яке потребує зберігання та обробки всіх векторів інвертованого списку для обчислення нового центроїда, рівняння (5.16) використовує лише один елемент – той, що наразі вставляється або видаляється, що відповідає стохастичній апроксимації повного субградієнта (4.15) його незміщеною оцінкою за одним елементом. Адаптивний варіант SQC-Adam автоматично підлаштовує крок оновлення для кожного центроїда окремо, оскільки центроїди розділів із високою частотою вставок отримують згладжені оновлення завдяки накопиченню квадратів субградієнтів у $v_{k^*}^t$, тоді як центроїди розділів із рідкісними модифікаціями зберігають достатню швидкість коригування. Ця властивість є особливо важливою для IVF індексів із дисбалансом розділів, де різні інвертовані списки оновлюються з суттєво різною частотою. Також послідовність оновлень (5.16) безпосередньо мінімізує похибку реконструкції $\bar{\varepsilon} = \frac{1}{N}\sum_{i=1}^{N} \|x_i - q_1(x_i)\|^2$, яка є пропорційною цільовій функції $F(y)$ згідно з (5.14). Збіжність цього процесу забезпечується теоремою 18: за умов виконання умов Роббінса-Монро на послідовність кроків ρ_t та компактності допустимої множини

позицій центроїдів, послідовність $\{c_1^{(t)}, \dots, c_{n_c}^{(t)}\}$ збігається до зв'язної компоненти множини стаціонарних точок задачі (5.14) з ймовірністю один [64]. Ця теоретична гарантія є суттєвою перевагою порівняно з евристичними підходами до оновлення індексу, де збіжність до оптимального розбиття не гарантується.

Водночас необхідно зазначити, що стохастичне оновлення центроїдів за правилом (5.16) не перерозподіляє вектори, вже присутні в інвертованих списках, між розділами. Призначення вектора x_i до інвертованого списку $\mathcal{J}_j$ залишається фіксованим з моменту вставки, навіть якщо зміщення центроїда c_j робить інший центроїд c_l ближчим до x_i. Для коригування цієї невідповідності може бути застосовано періодичне локальне перепризначення: для підмножини розділів із найбільшою похибкою реконструкції виконується перерозподіл векторів між сусідніми інвертованими списками на основі поточних позицій центроїдів. Критерій вибору розділів для локального перепризначення може базуватися на статистиках реального часу, таких як середня відстань між векторами та центроїдом розділу або коефіцієнт варіації розмірів інвертованих списків. Суттєвою перевагою є те, що такий локальний перерозподіл потребує перегляду лише невеликої підмножини сусідніх розділів, а не повного перебудування індексу, оскільки стохастичне оновлення центроїдів забезпечує поступову адаптацію глобальної структури розбиття до змін у розподілі векторів.

Зв'язок запропонованого підходу з інтеграцією PQ потребує окремого розгляду. У варіанті IVFADC [49] вектори зберігаються у стиснутому вигляді як коди квантизації добутку залишкових векторів $r_i = x_i - q_1(x_i)$. При зміщенні центроїда c_j за правилом (5.16) залишкові вектори усіх елементів інвертованого списку $\mathcal{J}_j$ формально змінюються, що потребувало б їхнього перекодування. Проте для малих кроків оновлення ρ_t, що гарантуються умовою $\sum_{t=0}^{\infty} \rho_t^2 < \infty$, зміщення центроїда на кожній ітерації є нехтовно малим порівняно з масштабом залишкових векторів, і перекодування може виконуватися ліниво – лише під час пошуку або за умови накопичення суттєвого зміщення центроїда, що перевищує заданий поріг $\|c_j^{(t)} - c_j^{(t_0)}\| > \delta$, де t_0 позначає ітерацію останнього перекодування.

У контексті IVF індексу кожна «ітерація» відповідає одній операції вставки або видалення, а «епоха» – пакету модифікацій, після якого виконується оцінювання якості індексу. Застосування усереднення по траєкторії (4.18) до

послідовності позицій центроїдів забезпечує згладжування стохастичних флуктуацій, що виникають при обробці окремих векторів, та стабілізує процес оновлення в умовах високої варіабельності вхідного потоку даних.

Застосування методу SQC для ітеративного оновлення інвертованого файлового індексу забезпечує три ключові переваги порівняно з існуючими підходами до оновлення індексу:

1. обчислювальна ефективність: оновлення центроїда виконується одночасно з операцією вставки без додаткових витрат на повне перерахування середнього за інвертованим списком, оскільки стохастична оцінка субградієнта (5.16) потребує лише одного вектора та має складність $O(d)$, що є краще порівняно зі складністю пошуку найближчого центроїда $O(n_c d)$;
2. теоретичні гарантії: збіжність послідовності центроїдів до стаціонарної точки задачі стохастичної квазіградієнтної кластеризації забезпечується теоремою 18 та результатами [64], на відміну від евристичних стратегій локального переіндексування, що не мають подібних гарантій;
3. адаптивність: варіант SQC-Adam автоматично підлаштовує швидкість оновлення для кожного центроїда на основі накопиченої градієнтної інформації, що є суттєвою перевагою для IVF індексів із нерівномірним розподілом операцій модифікації між розділами.

Таким чином, запропонований підхід реалізує інший шлях оновлення IVF індексу, що базується на неперервній стохастичній оптимізації замість дискретних подій перебудови, та забезпечує поступову адаптацію структури розбиття до змін у розподілі векторів без потреби у повній реконструкції, час якої для масивів масштабу мільярдів записів може сягати декількох діб обчислень.

5.4 Висновки до розділу

Отже, у цьому розділі продемонстровано прикладне застосування методу стохастичної квантизації SQC та його адаптивних модифікацій до трьох суттєво відмінних класів задач: кольорової квантизації зображень, класифікації високовимірних даних та оновлення індексних структур для наближеного пошуку найближчих сусідів. Для кожної задачі виконано формальне зведення до загальної

задачі стохастичної квантизації (4.12)-(4.13), що дозволило безпосередньо застосувати розроблені у попередніх розділах методи.

Чисельні експерименти на наборі даних ImageNet [108] підтвердили, що для задачі кольорової квантизації запропонований метод SQC скорочує кількість обчислень функції відстані приблизно на 70% порівняно з алгоритмом K-середніх [70] без втрати якості квантизації, що є перевагою для масштабованої обробки зображень високої роздільної здатності. Інтеграція методу SQC з архітектурою Triplet Network на наборі даних MNIST [67] забезпечила значення індексу Ранда $\mathrm{RI} = 0.9838$ при повній розмітці даних та зберегла високу узгодженість $\mathrm{RI} = 0.9774$ за умови лише 50% розмічених даних, що підтверджує ефективність комбінованого підходу для подолання «прокляття розмірності» у задачах класифікації. Для задачі оновлення інвертованого файлового індексу показано, що стохастичне правило оновлення центроїдів (5.16) забезпечує обчислювальну складність $O(d)$ на одну операцію оновлення з теоретичними гарантіями збіжності до стаціонарної точки задачі оптимальної квантизації.

На підставі отриманих результатів підтверджено практичну ефективність запропонованого методу стохастичної квантизації як універсального інструменту для широкого класу прикладних задач, що потребують масштабованого аналізу великих даних в умовах обмежених обчислювальних ресурсів та динамічних розподілів. Адаптивні модифікації SQC-Adam та SQC-RMSProp продемонстрували перевагу над базовим стохастичним субградієнтним спуском у режимах обмеженої розмітки даних та нерівномірного розподілу операцій модифікації, що мотивує подальше застосування комбінованого підходу до більш складних задач обробки високовимірних даних.

ВИСНОВКИ

Дисертаційну роботу присвячено дослідженню, розробці та впровадженню методів і моделей стохастичної квантизації для масштабованого кластерного аналізу великих даних. У роботі докладно розглянуто існуючі методи векторної квантизації та кластеризації на основі центроїдів і запропоновано новий підхід, який дає змогу адаптивно оцінювати оптимальні розміщення центроїдів за обмежених обчислювальних ресурсів та в умовах розподільного дрейфу, характерного для потоків даних.

Запропоноване рішення спирається на міждисциплінарні підходи з галузей стохастичної оптимізації, негладкого аналізу та машинного навчання «без учителя». Головним результатом роботи є створення методу стохастичної квазіградієнтної кластеризації (SQC), який зводить задачу оптимальної квантизації з рухомими центрами (4.3)-(4.4), сформульовану через метрику Канторовича-Рубінштейна (4.1) на відповідній транспортній задачі за обмежень (4.2), до задачі стохастичного програмування (4.12) із цільовим функціоналом (4.13) [64,76]. На основі методу SQC побудовано модифіковану структуру даних стохастичного інвертованого файлового індексу (SIVF), що забезпечує адаптивне розбиття простору ознак у векторних базах даних із теоретичними гарантіями збіжності та зменшеними вимогами до обчислювальних ресурсів.

У результаті виконання дисертаційної роботи вирішено такі наукові задачі:

1. Проведено порівняльний аналіз відомих підходів до векторної квантизації та кластеризації на основі центроїдів, включаючи класичний метод K-середніх та його модифікації – K-медіани (1.13), гармонічних K-середніх (1.16) та нечіткі K-середні (1.20) [70,110,117]. Виявлено ключові обмеження існуючих методів: чутливість до ініціалізації початкових наближень, висока обчислювальна складність повного оновлення центроїдів $O(Nd)$ на кожній ітерації, а також спільну проблему розподільного дрейфу (1.34) в умовах потоків даних [7,16];
2. На основі виявлених обмежень розроблено новий метод SQC, побудований на основі транспортної задачі з рухомими центрами [64] та стохастичної апроксимації Кіфера-Вольфовіца [55]. Стохастична оцінка субградієнта

(4.13) формується за правилом субдиференціювання функції мінімуму (теорема 15) як в одноелементному, так і в пакетному режимі (4.17). Завдяки використанню лише одного випадково обраного елемента $\tilde{\xi}$ замість повної сукупності $\{\xi_1, \cdots, \xi_N\}$, метод SQC зменшує вимоги до оперативної пам'яті обчислювального пристрою з $O(Nd)$ до $O(d)$, що є істотним для великомасштабних задач, де N може сягати порядку 10^4-10^6. Додатково представлено адаптивні модифікації, а також запропоновано детермінований метод ініціалізації центроїдів на основі жадібного алгоритму попереднього відбору кандидатів із задачею ЗЦЛП з обмеженнями кардинальності;

3. Виведено умови асимптотичної збіжності запропонованого методу SQC з використанням елементів негладкого аналізу. Доведення збіжності (теорема 18) ґрунтується на узагальненій диференційовності цільової функції (4.13) за теоремою 15 та на загальному результаті про збіжність узагальненого градієнтного методу (теорема 16) для негладких функцій із дисипативною стохастичною оцінкою градієнта. За умов Роббінса-Монро на послідовність кроків $\{\rho_t\}$ (2.41) та компактної опуклої допустимої множини Y встановлено, що з ймовірністю один послідовність наближень $\{y^t\}$ збігається до зв'язної компоненти множини стаціонарних точок Y^*. Цей результат поширюється і на адаптивну модифікацію SQC-Adam за рахунок збереження дисипативності оцінки субградієнта при покоординатній корекції кроку;
4. Запропоновано структуру даних SIVF – стохастичну модифікацію інвертованого файлового індексу IVF [7,49], призначену для адаптивного розбиття простору ознак у векторних базах даних. Встановлено еквівалентність задачі побудови індексу IVF (5.13) та задачі векторної квантизації методом SQC за порядку відстані $r = 2$ (5.14), що дозволяє замінити повний детерміністичний крок оновлення центроїдів (5.15) на ітеративне оновлення у режимі вставки та видалення елементів за правилом (5.16). Це забезпечує обчислювальну складність оновлення $O(d)$ замість $O(Nd)$ при додаванні нового вектора та можливість застосування адаптивного варіанту SQC-Adam для автоматичного балансування розмірів розділів індексу;

5. Проведено експериментальні дослідження запропонованого методу на відкритих наборах даних Iris [34], MNIST [67] та ImageNet [108], що підтвердило підвищення ефективності методів у порівнянні з існуючими рішеннями. На задачі кольорової квантизації зображень показано скорочення кількості обчислень функції відстані приблизно на 70% порівняно з алгоритмом K-середніх [70] без втрати якості квантизації за метрикою MSE (рисунок 5.1). На задачі семантичної кластеризації даних MNIST з використанням архітектури Triplet Network [42] зв'язка Triplet+SQC при повному наборі розміток досягла індексу Ранда $\mathrm{RI} = 0.9838$, перевершуючи базовий підхід Siamese+KNN ($\mathrm{RI} = 0.9790$), а адаптивні модифікації Triplet+SQC-RMSProp та Triplet+SQC-Adam продемонстрували найвищу стабільність в умовах обмеженої розмітки (таблиця 5.1).

Наукова новизна одержаних результатів полягає в тому, що:

1. Задачу оптимальної кластеризації векторних даних зведено до нової неопуклої негладкої задачі стохастичної оптимізації (4.12)-(4.13), що дозволяє для її розв'язання застосувати сучасні адаптивні методи стохастичної оптимізації, які використовуються для навчання DNN;
2. Розроблено новий метод стохастичної квазіградієнтної кластеризації, який відрізняється від існуючих методів векторної квантизації використанням субдиференційовної цільової функції транспортної відстані та стохастичної апроксимації, що дозволяє оцінювати оптимальні розміщення центроїдів лише з використанням підмножини навчальних даних, зменшуючи необхідну кількість обчислювальних ресурсів (оперативної пам'яті та обчислювального часу центрального процесора);
3. Запропоновано новий метод початкової ініціалізації центроїдів за допомогою жадібного алгоритму агрегації близьких даних та методів дискретної оптимізації, що покращує стійкість збіжності методу SQC порівняно зі стратегією K-середніх++ [5];
4. Удосконалено скінченно-різницевий метод оптимізації негладких неопуклих цільових функцій за допомогою методу згладжування (усереднення) функцій для пошуку точок екстремумів, використовуючи скінченно-різницеву апроксимацію градієнтів вздовж стохастичних напрямків. Отримано нові

умови асимптотичної збіжності запропонованого методу SQC (теорема 18) та нові оцінки швидкості збіжності пакетного стохастичного скінченно-різницевого методу оптимізації негладких опуклих функцій на основі операції згладжування за Стєкловим, які враховують як гіперпараметри задачі оптимізації (розмірність простору, оцінка діаметра області оптимізації, оцінка константи Ліпшиця), так і гіперпараметри алгоритму (початковий крок методу, число скінченних різниць);

5. Удосконалено структуру даних IVF для інкрементного індексування векторних даних, в якій для індексації використовується запропонований метод SQC в умовах динамічної зміни розподілу даних (сезонні зміни, соціологічні фактори тощо).

Практична значущість роботи підтверджується створенням та апробацією програмного забезпечення для адаптивного оновлення індексу IVF векторних баз даних в реальному часі. Методи та програмні компоненти, розроблені на базі описаної моделі, можуть бути використані для:

1. Інтеграції в системи керування векторними базами даних, такі як FAISS [48], для забезпечення сталої точності наближеного пошуку найближчих сусідів у динамічних середовищах без необхідності періодичної повторної індексації всієї бази даних;
2. Побудови пошукових рушіїв тексту та зображень за семантичною схожістю, а також систем підтримки прийняття рішень, основаних на великих даних, для більш точного кластерного аналізу та виявлення закономірностей на нерозмічених даних;
3. Стиснення зображень з втратами шляхом кольорової квантизації в RGB-просторі (5.1) та потокового опрацювання великих візуальних колекцій;
4. Розширення можливостей кластерного аналізу в умовах потоків даних, де дані безперервно збираються та індексуються протягом тривалих періодів часу, а статистичні властивості часто змінюються через сезонні коливання, соціологічні тенденції тощо.

Основні положення та висновки дисертаційного дослідження обговорювалися на наукових семінарах кафедри прикладної математики факультету прикладної

математики Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського» та отримали схвальні відгуки.

Результати дисертаційної роботи апробовані на низці міжнародних науково-практичних конференцій, зокрема: IV Міжнародній конференції «Чисельні обчислення: теорія та алгоритми (NUMTA 2023)» (м. Калабрія, Італія, 2023 р.) [85]; XVI та XVII Міжнародних конференціях «Освіта та дослідження в інформаційному суспільстві (ERIS 2023, ERIS 2025)» (м. Пловдів, Болгарія) [89,91]; XXVII Міжнародній конференції «АВТОМАТИКА 2024» (м. Дніпро, 2024 р.) [57]; XIV Міжнародній науково-практичній конференції «Сучасна кібернетика: Глушковські читання-2025» Інституту кібернетики імені В.М. Глушкова НАН України (м. Київ, 2024 р.) [59].

Дослідження виконані в дисертаційній роботі по проблематиці стохастичних методів оптимізації та кластерного аналізу великих даних, а також результати експериментів є складовою проекту Національного фонду досліджень України (НФДУ) № 2020.02/0121 «Аналітичні методи та машинне навчання в теорії керування і прийнятті рішень за умов конфлікту та невизначеності» (2020-2023~рр.), який виконувався на кафедрі прикладної математики факультету прикладної математики Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського».

Науково-практичну важливість результатів дисертації підтверджує впровадження результатів досліджень у навчальний процес на факультеті прикладної математики та навчально-наукового інституту прикладного системного аналізу Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського». Елементи розроблених методів та підходів включені до курсу «Сучасні методи оптимізації» в рамках аспірантської асистентської практики, що дозволяє студентам ознайомитися з сучасними методами кластерного аналізу основаних на теорії стохастичної оптимізації та елементах негладкого аналізу.

Незважаючи на досягнуті результати, запропонований метод SQC розглядає лише задачу векторної квантизації на основі центроїдів, що є окремим підкласом задач кластерного аналізу. Перспективним напрямком подальших досліджень є адаптація принципів стохастичної апроксимації до інших класів моделей

машинного навчання без учителя, зокрема до GMM, де стохастична оцінка параметрів сумішей - центрів μ_k, коваріаційних матриць Σ_k та вагових коефіцієнтів π_k – може забезпечити аналогічне зменшення вимог до оперативної пам'яті з $O(Nd + Kd^2)$ до $O(Kd^2)$ за рахунок одноелементних оновлень при збереженні ймовірнісної інтерпретації кластерної належності [117]. Іншим важливим напрямком є розширення стохастичного підходу на алгоритми кластеризації на основі оцінювання густини розподілу ймовірності, де адаптивне оновлення оцінок густини при надходженні нових спостережень $\xi_{t,i} \sim P_t$ з часозалежних розподілів (1.31) може вирішити проблему розподільного дрейфу для ширшого класу геометрій кластерів, включаючи кластери довільної форми [129].

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Abdi, H., Williams, L.J.: Principal component analysis. WIREs Computational Statistics 2(4), 433–459 (Jul 2010). https://doi.org/10.1002/wics.101
2. Achterberg, T.: Scip: solving constraint integer programs. Mathematical Programming Computation 1(1), 1–41 (Jul 2009). https://doi.org/10.1007/s12532-008-0001-1
3. Aloise, D., Deshpande, A., Hansen, P., Popat, P.: Np-hardness of euclidean sum-of-squares clustering. Machine Learning 75(2), 245–248 (Jan 2009). https://doi.org/10.1007/s10994-009-5103-0
4. Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., et al.: Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (Apr 2024). https://doi.org/10.1145/3620665.3640366
5. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. p. 1027–1035. SODA ’07, Society for Industrial and Applied Mathematics, USA (2007)
6. Balas, E., Ceria, S., Cornu´ejols, G.: Mixed 0-1 programming by lift-and-project in a branch-and-cut framework. Management Science 42(9), 1229–1246 (Sep 1996). https://doi.org/10.1287/mnsc.42.9.1229
7. Baranchuk, D., Douze, M., Upadhyay, Y., Yalniz, I.Z.: Dedrift: Robust similarity search under content drift (Oct 2023). https://doi.org/10.1109/iccv51070.2023.01012
8. Bayandina, A.S., Gasnikov, A.V., Lagunovskaya, A.A.: Gradient-free two-point methods for solving stochastic nonsmooth convex optimization problems with small non-random noises. Automation and Remote Control 79(8), 1399–1408 (2018). https://doi.org/10.1134/s0005117918080039
9. Beohar, D., Rasool, A.: Handwritten digit recognition of mnist dataset using deep learning state-of-the-art artificial neural network (ann) and convolutional neural

network (cnn). In: 2021 International Conference on Emerging Smart Computing and Informatics (ESCI). pp. 542–548 (2021). https://doi.org/10.1109/ESCI50559.2021.9396870

10. Berahas, A.S., Cao, L., Choromanski, K., Scheinberg, K.: A theoretical and empirical comparison of gradient approximations in derivative-free optimization (2021)
11. Bertsekas, D.P., Nedic, A., Ozdaglar, A.E.: Convex analysis and optimization. Athena Scientific (2003)
12. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Springer US (1981). https://doi.org/10.1007/978-1-4757-0450-1
13. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., Brock, A., Paganini, M., Irving, G., Vinyals, O., Osindero, S., Simonyan, K., Rae, J.W., Elsen, E., Sifre, L.: Improving language models by retrieving from trillions of tokens (2022), https://arxiv.org/abs/2112.04426
14. Bottou, L.: Large-scale machine learning with stochastic gradient descent (2010). https://doi.org/10.1007/978-3-7908-2604-3_16
15. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM Review 60(2), 223–311 (2018). https://doi.org/10.1137/16m1080173
16. Cao, X., Zhang, J., Poor, H.V.: On the time-varying distributions of online stochastic optimization (Jul 2019). https://doi.org/10.23919/acc.2019.8814889
17. Celebi, M.E.: Improving the performance of k-means for color quantization. Image and Vision Computing 29(4), 260–271 (Mar 2011). https://doi.org/10.1016/j.imavis.2010.10.002
18. Celebi, M.E., Kingravi, H.A., Vela, P.A.: A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Systems with Applications 40(1), 200–210 (Jan 2013). https://doi.org/10.1016/j.eswa.2012.07.021

19. Cesa-Bianchi, N., Conconi, A., Gentile, C.: On the generalization ability of on-line learning algorithms. IEEE Transactions on Information Theory 50(9), 2050–2057 (sept 2004). https://doi.org/10.1109/tit.2004.833339
20. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection. ACM Computing Surveys 41(3), 1–58 (Jul 2009). https://doi.org/10.1145/1541880.1541882
21. Cuesta-Albertos, J.A., Gordaliza, A., Matr´an, C.: Trimmed k-means: an attempt to robustify quantizers. The Annals of Statistics 25(2) (Apr 1997). https://doi.org/10.1214/aos/1031833664
22. Dakin, R.J.: A tree-search algorithm for mixed integer programming problems. The computer journal 8(3), 250–255 (1965)
23. Dantzig, G.: Linear Programming and Extensions. Princeton University Press (Dec 1963). https://doi.org/10.1515/9781400884179
24. Deisenroth, M.P., Ong, C.S., Faisal, A.A.: Mathematics for Machine Learning. Cambridge University Press (2021)
25. Dhillon, I.S., Guan, Y., Kulis, B.: Kernel k-means (Aug 2004). https://doi.org/10.1145/1014052.1014118
26. Dorigo, M., Blum, C.: Ant colony optimization theory: A survey. Theoretical Computer Science 344(2–3), 243–278 (Nov 2005). https://doi.org/10.1016/j.tcs.2005.05.020
27. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. 12(null), 2121–2159 (Jul 2011). https://doi.org/10.5555/1953048.2021068
28. Duchi, J.C., Jordan, M.I., Wainwright, M.J., Wibisono, A.: Optimal rates for zero-order convex optimization: The power of two function evaluations. IEEE Transactions on Information Theory 61(5), 2788–2806 (2015). https://doi.org/10.1109/tit.2015.2409256
29. Dunn, J.C.: A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. Journal of Cybernetics 3(3), 32–57 (Jan 1973). https://doi.org/10.1080/01969727308546046
30. E. Alpaydin, C.K.: Optical recognition of handwritten digits (1998). https://doi.org/10.24432/C50P49

31. Eremin, I.I.: The penalty method in convex programming. Cybernetics 3(4), 53–56 (1971). https://doi.org/10.1007/bf01071708
32. Ermoliev, Y.: Stochastic Programming Methods. Nauka, Moscow (1976)
33. Ermoliev, Y.M., Norkin, V.I., Wets, R.J.B.: The minimization of semicontinuous functions: Mollifier subgradients. SIAM Journal on Control and Optimization 33(1), 149–167 (1995). https://doi.org/10.1137/s0363012992238369
34. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of Eugenics 7(2), 179–188 (Sep 1936). https://doi.org/10.1111/j.1469-1809.1936.tb02137.x
35. Forgy, E.W.: Cluster analysis of multivariate data: Efficiency versus interpretability of classification. Biometrics 21(3), 768–780 (1965)
36. Garc´ıa-Escudero, L.A., Gordaliza, A., Matr´an, C., Mayo-Iscar, A.: A general trimming approach to robust cluster analysis. The Annals of Statistics 36(3) (Jun 2008). https://doi.org/10.1214/07-aos515
37. Gomory, R.E.: An algorithm for the mixed integer problem. RAND Corporation, Santa Monica, CA (1960)
38. Groll, L., Jakel, J.: A new convergence proof of fuzzy c-means. IEEE Transactions on Fuzzy Systems 13(5), 717–720 (Oct 2005). https://doi.org/10.1109/tfuzz.2005.856560
39. Gupal, A.M.: A method for the minimization of almost-differentiable functions. Cybernetics 13(1), 115–117 (1977)
40. Hamerly, G., Elkan, C.: Alternatives to the k-means algorithm that find better clusterings (Nov 2002). https://doi.org/10.1145/584792.584890
41. 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 R´io, J.F., Wiebe, M., Peterson, P., G´erard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E.: Array programming with NumPy. Nature 585(7825), 357–362 (Sep 2020). https://doi.org/10.1038/s41586-020-2649-2
42. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: Feragen, A., Pelillo, M., Loog, M. (eds.) Similarity-Based Pattern Recognition. pp. 84–92. Springer International Publishing, Cham (2015)

43. Hu, Z., Xiong, S., Su, Q., Zhang, X.: Sufficient conditions for global convergence of differential evolution algorithm. Journal of Applied Mathematics 2013, 1–14 (2013). https://doi.org/10.1155/2013/193196
44. Huang, L., Chen, X., Huo, W., Wang, J., Zhang, F., Bai, B., Shi, L.: Branch and bound in mixed integer linear programming problems: A survey of techniques and trends (2021), https://arxiv.org/abs/2111.06257
45. Hunter, J.D.: Matplotlib: A 2d graphics environment. Computing in Science & Engineering 9(3), 90–95 (2007). https://doi.org/10.1109/MCSE.2007.55
46. Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern Recognition Letters 31(8), 651–666 (Jun 2010). https://doi.org/10.1016/j.patrec.2009.09.011
47. Jin, C., Zhang, Y., Balakrishnan, S., Wainwright, M.J., Jordan, M.I.: Local maxima in the likelihood of gaussian mixture models: structural results and algorithmic consequences. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. p. 4123–4131. NIPS'16, Curran Associates Inc., Red Hook, NY, USA (2016)
48. Johnson, J., Douze, M., Jegou, H.: Billion-scale similarity search with gpus. IEEE Transactions on Big Data 7(3), 535–547 (Jul 2021). https://doi.org/10.1109/tbdata.2019.2921572
49. J´egou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(1), 117–128 (2011). https://doi.org/10.1109/TPAMI.2010.57
50. Karaboga, D., Basturk, B.: Artificial bee colony (abc) optimization algorithm for solving constrained optimization problems. https://doi.org/10.1007/978-3-540-72950-1_77
51. Karloff, H.: The simplex algorithm (2009). https://doi.org/10.1007/978-0-8176-4844-2_2
52. Kennedy, J., Eberhart, R.: Particle swarm optimization. https://doi.org/10.1109/icnn.1995.488968
53. Kerr, G., Ruskin, H., Crane, M., Doolan, P.: Techniques for clustering gene expression data. Computers in Biology and Medicine 38(3), 283–293 (Mar 2008). https://doi.org/10.1016/j.compbiomed.2007.11.001

54. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS ’20, Curran Associates Inc., Red Hook, NY, USA (2020)
55. Kiefer, J., Wolfowitz, J.: Stochastic estimation of the maximum of a regression function. The Annals of Mathematical Statistics 23(3), 462–466 (Sep 1952). https://doi.org/10.1214/aoms/1177729392
56. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017), https://doi.org/10.48550/arXiv.1412.6980
57. Kozyriev, A., Norkin, V.: Stysnennia zobrazhennia z vtratyamy za dopomohoiu stokhastychnoho kvantuvannia (2024), http://automatika2024.dp.ua/
58. Kozyriev, A.: kaydotdev/stochastic-quantization: Robust and scalable stochastic quasigradient clustering (Aug 2024), https://github.com/kaydotdev/stochastic-quantization
59. Kozyriev, A., Norkin, V.: Stochastychnyi invertovanyi filovyi index (2025), http://glushkov.kpi.ua/, https://drive.google.com/file/d/1PTyvCYtyEEZjm7DUGsd07NUJeqqOgb
60. Kriegel, H.P., Kr¨oger, P., Zimek, A.: Clustering high-dimensional data. ACM Transactions on Knowledge Discovery from Data 3(1), 1–58 (Mar 2009). https://doi.org/10.1145/1497577.1497578
61. Krishna, K., Ramakrishnan, K., Thathachar, M.: Vector quantization using genetic k-means algorithm for image compression. In: Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat. vol. 3, pp. 1585–1587 vol.3 (1997). https://doi.org/10.1109/ICICS.1997.652261
62. Krishnaswamy, R., Li, S., Sandeep, S.: Constant approximation for k-median and k-means with outliers via iterative rounding (Jun 2018). https://doi.org/10.1145/3188745.3188882
63. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C., Bottou, L., Weinberger,

K. (eds.) Advances in Neural Information Processing Systems. vol. 25. Curran Associates, Inc. (2012)

64. Kuzmenko, V., Uryasev, S.: Kantorovich–rubinstein distance minimization: Application to location problems. Springer Optimization and Its Applications 149, 59–68 (Sep 2019). https://doi.org/10.1007/978-3-030-22788-3_3

65. Lakshmanan, R., Pichler, A.: Soft quantization using entropic regularization. Entropy 25(10), 1435 (Oct 2023). https://doi.org/10.3390/e25101435

66. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998). https://doi.org/10.1109/5.726791

67. LeCun, Y., Cortes, C., Burges, C.: Mnist handwritten digit database. ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist 2 (2010)

68. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., K¨uttler, H., Lewis, M., Yih, W.t., Rockt¨aschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)

69. Linderoth, J.T., Savelsbergh, M.W.P.: A computational study of search strategies for mixed integer programming. INFORMS Journal on Computing 11(2), 173–187 (May 1999). https://doi.org/10.1287/ijoc.11.2.173

70. Lloyd, S.: Least squares quantization in pcm. IEEE Transactions on Information Theory 28(2), 129–137 (Mar 1982). https://doi.org/10.1109/tit.1982.1056489

71. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. vol. 5, pp. 281–298. University of California press (1967)

72. Mahajan, M., Nimbhorkar, P., Varadarajan, K.: The planar k-means problem is np-hard (2009). https://doi.org/10.1007/978-3-642-00202-1_24

73. Marchand, H., Martin, A., Weismantel, R., Wolsey, L.: Cutting planes in integer and mixed integer programming. Discrete Applied Mathematics 123(1–3), 397–446 (Nov 2002). https://doi.org/10.1016/s0166-218x(01)00348-1

74. Mashtalir, S.V., Stolbovyi, M.I., Yakovlev, S.V.: Clustering video sequences by the method of harmonic k-means. Cybernetics and Systems Analysis 55(2), 200–206 (Mar 2019). https://doi.org/10.1007/s10559-019-00124-9
75. Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N., Capstick, E., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J.C., Shoham, Y., Wald, R., Walsh, T., Hamrah, A., Santarlasci, L., Lotufo, J.B., Rome, A., Shi, A., Oak, S.: Artificial intelligence index report 2025 (2025), https://arxiv.org/abs/2504.07139
76. Mikhalevich, V.S., Gupal, A.M., Norkin, V.I.: Methods of nonconvex optimization (2024), https://arxiv.org/abs/2406.10406
77. Montavon, G., Orr, G., M¨uller, K.R. (eds.): Neural networks: Tricks of the trade. Springer (2012)
78. Murasaki, K., Ando, S., Shimamura, J.: Semi-supervised representation learning via triplet loss based on explicit class ratio of unlabeled data. IEICE Transactions on Information and Systems E105.D(4), 778–784 (Apr 2022). https://doi.org/10.1587/transinf.2021edp7073
79. Murray, A., van Kemenade, H., wiredfool, Clark, J.A., Karpinsky, A., Baranoviˇc, O., Gohlke, C., Yay295, Dufresne, J., Brett, M., DWesl, Schmidt, D., Kopachev, K., Houghton, A., REDxEYE, Mani, S., Landey, S., Koskela, A., Ware, J., vashek, Piolie, Douglas, J., T., S., Caro, D., Martinez, U., Kossouho, S., Lahd, R.: python-pillow/pillow: 10.4.0 (Jul 2024). https://doi.org/10.5281/zenodo.12606429
80. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM Journal on Optimization 19(4), 1574–1609 (2009). https://doi.org/10.1137/070704277
81. Nesterov, Y.E.: A method of solving a convex programming problem with convergence rate $O(1/k^2)$. In: Doklady Akademii Nauk. vol. 269, pp. 543–547. Russian Academy of Sciences (1983)
82. Nesterov, Y.E.: Introductory lectures on convex optimization: A basic course, vol. 103. Kluwer Academic Publishers (2004). https://doi.org/10.1007/s10107-004-0552-5

83. Newton, D., Yousefian, F., Pasupathy, R.: Stochastic gradient descent: Recent trends. Recent Advances in Optimization and Modeling of Contemporary Problems p. 193–220 (2018). https://doi.org/10.1287/educ.2018.0191
84. Nguyen, C.D., Duong, T.H.: K-means** – a fast and efficient k-means algorithms. International Journal of Intelligent Information and Database Systems 11(1), 27 (2018). https://doi.org/10.1504/ijiids.2018.091595
85. Norkin, V., Pichler, A., Kozyriev, A.: Constrained global optimization by smoothing (2023), https://www.numta.org/pdf/NUMTA2023%5FBook.pdf
86. Norkin, V.I.: Stochastic generalized gradient methods for training nonconvex nonsmooth neural networks. Cybernetics and Systems Analysis 57(5), 714–729 (2021). https://doi.org/10.1007/s10559-021-00397-z
87. Norkin, V.: Two random search algorithms for minimizing non-differentiable functions. In: Ermoliev, Y.M., Kovalenko, I.N. (eds.) Mathematical Methods of Operations Research and Reliability Theory, pp. 36–40. Institute of Cybernetics of the NAS of Ukraine, Kyiv (1978)
88. Norkin, V.: A stochastic smoothing method for nonsmooth global optimization. Cybernetics and Computer Technologies 1, 5–14 (2020). https://doi.org/10.34229/2707-451x.20.1.1
89. Norkin, V., Kozyriev, A.: Modern quasigradient stochastic algorithms (2023), https://eris.adis.org/wp-content/uploads/2023/10/ERIS-Program%5F2023.pdf
90. Norkin, V., Kozyriev, A.: On shor's r-algorithm for problems with constraints. Cybernetics and Computer Technologies p. 16–22 (sept 2023). https://doi.org/10.34229/2707-451x.23.3.2
91. Norkin, V., Kozyriev, A.: K-means seeding using mixed-integer optimization (2025), https://eris.adis.org/wp-content/uploads/2025/09/ERIS-2025-Preliminary-programme-v4.pdf
92. Norkin, V., Kozyriev, A., Norkin, B.: Modern stochastic quasigradient optimization algorithms. International Scientific Technical Journal "Problems of Control and Informatics" 69(2), 71–83 (Mar 2024). https://doi.org/10.34229/1028-0979-2024-2-6
93. Norkin, V., Pichler, A., Kozyriev, A.: Constrained global optimization by smoothing (2025). https://doi.org/10.1007/978-3-031-81241-5_10

94. Orchard, M., Bouman, C.: Color quantization of images. IEEE Transactions on Signal Processing 39(12), 2677–2690 (1991). https://doi.org/10.1109/78.107417
95. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
96. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation (2014). https://doi.org/10.3115/v1/d14-1162
97. Perron, L., Furnon, V.: Or-tools, https://developers.google.com/optimization/
98. Poliak, B.T.: The Heavy Ball Method, p. 65–68. Optimization Software (1987)
99. Polyak, B.: Some methods of speeding up the convergence of iteration methods. USSR Computational Mathematics and Mathematical Physics 4(5), 1–17 (1964). https://doi.org/10.1016/0041-5553(64)90137-5
100. Qian, X., Klabjan, D.: The impact of the mini-batch size on the variance of gradients in stochastic gradient descent (2020), https://arxiv.org/abs/2004.13146
101. Radomirovi´c, B., Jovanovi´c, V., Nikoli´c, B., Stojanovi´c, S., Venkatachalam, K., Zivkovic, M., Njeguˇs, A., Bacanin, N., Strumberger, I.: Text document clustering approach by improved sine cosine algorithm. Information Technology and Control 52(2), 541–561 (Jul 2023). https://doi.org/10.5755/j01.itc.52.2.33536
102. Rand, W.M.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66(336), 846–850 (Dec 1971). https://doi.org/10.1080/01621459.1971.10482356
103. Reddy, C.K., Vinzamuri, B.: A survey of partitional and hierarchical clustering algorithms (sept 2018). https://doi.org/10.1201/9781315373515-4
104. Robbins, H., Monro, S.: A stochastic approximation method. The Annals of Mathematical Statistics 22(3), 400–407 (Sep 1951). https://doi.org/10.1214/aoms/1177729586
105. Rockafellar, R.T., Wets, R.J.B.: Variational Analysis. Springer Berlin Heidelberg (1998). https://doi.org/10.1007/978-3-642-02431-3
106. Ruder, S.: An overview of gradient descent optimization algorithms (2017), https://arxiv.org/abs/1609.04747

107. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature 323(6088), 533–536 (Oct 1986). https://doi.org/10.1038/323533a0

108. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision (IJCV) 115(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y

109. Scavuzzo, L., Aardal, K., Lodi, A., Yorke-Smith, N.: Machine learning augmented branch and bound for mixed integer linear programming. Mathematical Programming (Aug 2024). https://doi.org/10.1007/s10107-024-02130-y

110. Scitovski, R., Sabo, K., Mart´inez- ´Alvarez, F., Ungar, ˇS.: Data Clustering, pp. 31–64. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-74552-3_3

111. Sculley, D.: Web-scale k-means clustering. Proceedings of the 19th international conference on World wide web p. 1177–1178 (Apr 2010). https://doi.org/10.1145/1772690.1772862

112. Shamir, O.: An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. The Journal of Machine Learning Research 18(1), 1703–1713 (2017)

113. Shapiro, A., Dentcheva, D., Ruszczy´nski, A.: Lectures on Stochastic Programming. Society for Industrial and Applied Mathematics (2009). https://doi.org/10.1137/1.9780898718751

114. Sheather, S.J.: Density estimation. Statistical Science 19(4) (Nov 2004). https://doi.org/10.1214/088342304000000297

115. Spielman, D.A., Teng, S.H.: Smoothed analysis of algorithms. Journal of the ACM 51(3), 385–463 (May 2004). https://doi.org/10.1145/990308.990310

116. Stefan Aeberhard, M.F.: Wine (1992). https://doi.org/10.24432/C5PC7J

117. Steinley, D.: K-means clustering: A half-century synthesis. British Journal of Mathematical and Statistical Psychology 59(1), 1–34 (May 2006). https://doi.org/10.1348/000711005x48266

118. Storn, R., Price, K.: Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. Journal of Global Optimization 11(4), 341–359 (Dec 1997). https://doi.org/10.1023/A:1008202821328

119. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: Dasgupta, S., McAllester, D. (eds.) Proceedings of the 30th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 28, pp. 1139–1147. PMLR, Atlanta, Georgia, USA (17–19 Jun 2013)

120. Tabianan, K., Velu, S., Ravi, V.: K-means clustering approach for intelligent customer segmentation using customer purchase behavior data. Sustainability 14(12), 7243 (Jun 2022). https://doi.org/10.3390/su14127243

121. Tieleman, T., Hinton, G.: Rmsprop: Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. COURSERA Neural Networks Mach. Learn 17 (2012)

122. Turpault, N., Serizel, R., Vincent, E.: Semi-supervised triplet loss based learning of ambient audio embeddings. ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (May 2019). https://doi.org/10.1109/icassp.2019.8683774

123. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6000–6010. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)

124. Walkington, N.J.: Nesterov's method for convex optimization. SIAM Review 65(2), 539–562 (2023). https://doi.org/10.1137/21M1390037

125. Wan, X.: Application of k-means algorithm in image compression. IOP Conference Series: Materials Science and Engineering 563(5) (Jul 2019). https://doi.org/10.1088/1757-899x/563/5/052042

126. Wang, W., Zhang, Y.: On fuzzy cluster validity indices. Fuzzy Sets and Systems 158(19), 2095–2117 (Oct 2007). https://doi.org/10.1016/j.fss.2007.03.004

127. Widodo, A., Budi, I.: Clustering patent document in the field of ict (information & communication technology). In: 2011 International Conference on

Semantic Technology and Information Retrieval. pp. 203–208 (2011). https://doi.org/10.1109/STAIR.2011.5995789

128. Wolsey, L.A.: Integer Programming. Wiley (2021)

129. Xu, R., Wunsch, D.: Survey of clustering algorithms. IEEE Transactions on Neural Networks 16(3), 645–678 (2005). https://doi.org/10.1109/TNN.2005.845141

130. Xuan, H., Stylianou, A., Pless, R.: Improved embeddings with easy positive triplet mining (2020), https://arxiv.org/abs/1904.04370

131. Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity Search The Metric Space Approach. Springer US (2006). https://doi.org/10.1007/0-387-29151-2

132. Zhang, B., Hsu, M., Dayal, U.: K-harmonic means – a spatial clustering algorithm with boosting (2001). https://doi.org/10.1007/3-540-45244-3_4

133. Zorich, V.A.: Mathematical Analysis I. Springer Berlin Heidelberg (2015). https://doi.org/10.1007/978-3-662-48792-1