Голові
разової спеціалізованої вченої ради
Інституту математики НАН України
доктору фізико-математичних наук,
заступнику директора з наукових питань
Інституту математики НАН України
Василику Віталію Богдановичу

РЕЦЕНЗІЯ

на дисертаційну роботу **Шамрая Максима Борисовича** на тему:
“Low-rank approximations, perturbation bounds, and their role in model compression”
(“Низькорангові наближення, оцінки збурень та їх роль у компресії моделей”),
подану на здобуття ступеня доктора філософії
у галузі знань 11 Математика та статистика
за спеціальністю 113 Прикладна математика

У дисертаційній роботі Максима Борисовича Шамрая представлено низку нових вагомих результатів, пов’язаних із розробленням математично обґрунтованих методів компресії моделей машинного навчання на основі низькорангових наближень, теорії збурень та аналізу структурних змін параметрів. Робота присвячена актуальній науковій проблемі прикладної математики — побудові єдиного підходу до компресії великих моделей із гарантованим контролем похибки та якості розв’язання задач, пов'язаних із математичними моделями нейронних мереж. Актуальність теми обумовлена стрімким зростанням розмірності сучасних нейронних моделей, що обмежує їх практичне використання через вимоги до пам’яті, швидкодії та енергоспоживання, тоді як існуючі методи компресії переважно мають евристичний характер.

Можна виділити такі основні наукові результат даної дисертаційної роботи:

- Встановлено нові оцінки збурень сингулярних значень для конкатенованих матриць, які дозволяють кількісно аналізувати вплив блокових змін на спектральні характеристики та якість низькорангових наближень.
- Виконано формалізацію задачі компресійно-орієнтованого групування матриць із використанням критерію допустимої похибки, що забезпечує перехід від евристичних підходів до строгих алгоритмічних рішень.

- Знайдено глобальні та інкрементні оцінки похибки реконструкції, які дозволяють здійснювати послідовне об’єднання матриць із гарантованим контролем якості.
- Побудовано три ефективні стратегії кластеризації для задачі спільної компресії, що відрізняються рівнем точності та обчислювальною складністю.
- Показано обмеженість класичних методів кластеризації для задач компресії та обґрунтовано необхідність спеціалізованих підходів, орієнтованих на контроль похибки.
- Виконано детальний аналіз сучасних методів прунінгу великих мовних моделей та встановлено вплив мовної специфіки на результати компресії.
- Знайдено новий підхід до побудови геометрії мов, який базується на аналізі структурних властивостей ваг нейронних мереж.
- Показано, що отриманий метричний простір мов відображає змістовні мовні зв’язки та може бути використаний для подальших досліджень у галузі обробки природної мови.
- Одержано аналітичні оцінки впливу прунінгу на якість нейромережевих політик у задачах керування та навчання з підкріпленням.
- Показано практичну ефективність запропонованих методів на реальних задачах компресії моделей, що підтверджує їх прикладну цінність .

Наукова новизна отриманих результатів полягає у створенні єдиного узгодженого підходу, який поєднує теорію збурень, низькорангові наближення, компресійно-орієнтовану кластеризацію та аналіз прунінгу великих моделей. Практичне значення роботи полягає у можливості застосування запропонованих методів для побудови ефективних алгоритмів компресії з контрольованою похибкою у широкому спектрі задач.

Дисертація має чітку структуру, логічно побудовану від огляду сучасного стану проблеми до отримання нових теоретичних результатів і їх практичної перевірки. У першому розділі проведено ґрунтовний аналіз літератури: від базових засад побудови нейронних мереж, сучасних архітектур, теорії оптимального керування та навчання з підкріпленням до сучасних методів компресії моделей. Основну увагу приділено підходам, заснованим на факторизації матриць і прунінгу. Наведено фундаментальні результати теорії збурень та низькорангових наближень, що слугують теоретичною основою подальших досліджень. Сформульовано задачу низькорангового наближення конкатенованих матриць та розглянуто підходи до її розв’язання, зокрема інкрементальні методи SVD. Окрему увагу приділено постановці задачі пошарового прунінгу нейронних мереж; зокрема розглянуто класичну задачу

«optimal brain damage», яка є фундаментальною для сучасних методів структурної компресії. Таким чином, цей розділ узагальнює відомі результати й формує чітке підґрунтя для постановки нових задач, підкреслюючи обмеження існуючих підходів.

Другий розділ присвячено дослідженню оцінок похибки низькорангового наближення та збурень сингулярних значень конкатенованих матриць. Зокрема, у теоремі 2.5 отримано нетривіальну оцінку збурення сингулярних значень, що виражається через операторні норми блоків матриці, і яка враховує її конкатеновану структуру. У цьому ж розділі запропоновано жадібні алгоритми кластеризації матриць, що базуються на контролі похибки низькорангового наближення їх конкатенації. Теоретичне обґрунтування цих алгоритмів спирається на отримані оцінки: зокрема, у теоремі 2.11 та наслідку 2.15 виведено верхні оцінки похибки низькорангового наближення, які мають нестрогий характер, тоді як у наслідку 2.18 запропоновано наближення цієї похибки, придатне для використання в алгоритмічних процедурах. Сукупність цих результатів дозволяє сформулювати критерій допустимості об’єднання матриць у кластер із контролем якості наближення. Важливість отриманих результатів полягає в тому, що вони вперше забезпечують узгоджене поєднання спектральної теорії та алгоритмічних процедур кластеризації, що дозволяє переходити від евристичних методів до підходів із гарантованим контролем похибки, що є критично важливим для застосувань у великих моделях машинного навчання.

У третьому розділі фокус дослідження зміщується на задачі прунінгу нейронних мереж. Розділ починається з чисельних експериментів, проведених для великих мовних моделей, у яких продемонстровано, що ефективність прунінгу суттєво залежить від мови, на якій здійснюється калібрування. На основі цих результатів побудовано метричний простір мов, визначений через сигнали важливості параметрів моделі. У подальшій частині розділу наведено теоретичні результати, що встановлюють оцінки зверху для якості нейронних контролерів після прунінгу. Зокрема, у теоремі 3.10 отримано оцінку погіршення якості стратегії керування з урахуванням структурних змін параметрів, а у теоремі 3.19 ці результати узгоджено з лемою про різницю продуктивності (лема 3.18), що дозволяє отримати узагальнену оцінку впливу прунінгу на функціонал якості. Отримані результати мають суттєве значення, оскільки забезпечують теоретичне підґрунтя для застосування прунінгу у відповідальних задачах, де необхідно гарантувати збереження якості, зокрема в задачах керування та прийняття рішень, і водночас відкривають нові перспективи для аналізу внутрішньої структури моделей машинного навчання.

Отримані результати є взаємопов’язаними та утворюють цілісну наукову концепцію. Практичне значення роботи полягає у можливості застосування отриманих результатів для побудови ефективних алгоритмів компресії моделей із контрольованою похибкою, що є важливим для широкого кола задач — від обробки природної мови до систем керування складними динамічними об’єктами.

Дисертацію оформлено відповідно до вимог, що висуваються до кваліфікаційних робіт на здобуття ступеня доктора філософії. Порушень академічної доброчесності у дисертації та наукових працях М.Б. Шамрая не виявлено.

Зауваження та побажання до дисертаційної роботи:

- Дисертаційна робота є комплексним і багатоплановим дослідженням, однак у деяких місцях виклад матеріалу є досить щільним, що ускладнює сприйняття окремих результатів, особливо для читачів, які не є фахівцями у відповідній галузі. Окремі позначення та терміни могли б бути введені більш поступово з метою покращення структури тексту.
- У другому розділі наведено детальне доведення твердження 2.2, що стосується оцінки норми блокової матриці. Разом з тим, цей результат має допоміжний характер і є достатньо очевидним з огляду на відомі властивості операторних норм, тому його повне доведення видається надмірним; доцільним могло б бути посилання на відповідні відомі результати або наведення короткого обґрунтування.
- Поряд із ґрунтовними теоретичними результатами щодо впливу прунінгу на якість нейронних контролерів, доцільно було б розширити експериментальну частину шляхом проведення чисельних досліджень саме для задач керування. Зокрема, це дозволило б емпірично підтвердити отримані оцінки впливу прунінгу на поведінку контролера та відповідні показники якості, що підсилить практичну значущість наведених теоретичних результатів.

Разом з тим зазначені зауваження не знижують загальної високої оцінки дисертаційної роботи. Дисертація є завершеним науковим дослідженням, виконаним на високому теоретичному та прикладному рівнях, а її результати мають істотне значення для розвитку прикладної математики та сучасного машинного навчання.

Дисертаційна робота і наукові публікації Максима Борисовича Шамрая відповідають спеціальності 113 Прикладна математика та задовольняють вимоги постанови № 44 Кабінету Міністрів України від 12 січня 2022 р. «Про затвердження Порядку присудження ступеня доктора філософії та скасування рішення разової спеціалізованої вченої ради закладу вищої освіти, наукової установи про

присудження ступеня доктора філософії», а їх автор заслуговує на присудження ступеня доктора філософії.

Рецензент:
професор, завідувач відділу
математичних проблем механіки та теорії керування
Інституту математики НАН України,
член-кореспондент НАН України,
доктор фізико-математичних наук,

Мазко Олексій Григорович

21 квітня 2026 р.