

На правах рукописи

УДК 519.853



Аникин Антон Сергеевич

**Параллельные вычислительные технологии решения
конечномерных задач оптимизации большой
размерности**

Специальность 05.13.01 —

«Системный анализ, управление и обработка информации»

АВТОРЕФЕРАТ

Диссертация на соискание учёной степени
кандидата физико-математических наук

Иркутск – 2020

Работа выполнена в Федеральном государственном бюджетном учреждении науки «Институт динамики систем и теории управления им. В.М. Матросова Сибирского отделения Российской академии наук».

Научный руководитель: Горнов Александр Юрьевич, доктор технических наук.

Ведущая организация: Федеральное государственное бюджетное учреждение науки «Институт математики и механики им. Н.Н. Красовского Уральского отделения Российской академии наук».

Защита состоится «28» декабря 2020 г. в 14:00 на заседании диссертационного совета ФПМИ.05.13.01.002. по адресу: 141701, Московская область, г. Долгопрудный, Институтский переулок, д. 9, МФТИ.

С диссертацией можно ознакомиться в библиотеке и на сайте Федерального государственного автономного образовательного учреждения высшего образования «Московский физико-технический институт (национальный исследовательский университет)»: <https://mipt.ru/education/post-graduate/soiskateli-fiziko-matematicheskie-nauki.php>

Работа представлена «26» сентября 2020 г. в Аттестационную комиссию федерального государственного автономного образовательного учреждения высшего образования «Московский физико-технический институт (национальный исследовательский университет)» для рассмотрения советом по защите диссертаций на соискание учёной степени кандидата наук в соответствии с п. 3.1 ст. 4 Федерального закона «О науке и государственной научно-технической политике».

Актуальность. Задачи оптимизации, безусловно, следует отнести к одной из важнейших проблем современной прикладной математики. Вместе с тем, они естественным образом возникают в нашей повседневной жизни, что связано с вполне объяснимым человеческим желанием максимизировать эффективность своей деятельности. Часть таких ежедневных, несложных, проблем, как, например, выбор оптимального маршрута поездки, или минимизация затрат на покупку требуемых товаров, в ряде случаев успешно решается людьми с использованием накопленного ими ранее жизненного опыта, причём, зачастую это выполняется «на интуитивном уровне», без особых раздумий. Решение же сложных оптимизационных задач без привлечения соответствующего математического аппарата и накопленной алгоритмической базы представляется малореальным. Сложные постановки экстремальных задач присущи широкому кругу прикладных областей, многие практически значимые проблемы могут быть сведены к строго формализованному виду.

За свою достаточно долгую историю математическая теория оптимизации прошла путь от простейших методов решения задач, которые в наше время уже считаются тривиальными, до современных, достаточно сложных, комплексных подходов, позволяющих успешно работать с постановками, которые ранее воспринимались не иначе, как неразрешимые. С формальной позиции многие задачи оптимизации можно разбить на две основные группы: конечномерная (параметрическая) оптимизация и оптимизация в бесконечномерном пространстве. К последней группе в первую очередь стоит отнести задачи вариационного исчисления и задачи поиска оптимального управления. Диссертационная работа посвящена рассмотрению современных задач из первой группы — задач конечномерной оптимизации большой размерности. Такие задачи, в свою очередь, могут быть разделены на два традиционных класса: задачи локальной оптимизации и задачи поиска глобального экстремума. И если для решения задач локальной оптимизации к настоящему времени разработаны теоретически обоснованные методы и подходы, то для задач поиска глобального экстремума ситуация гораздо более сложная. С сожалением приходится констатировать, что все детерминированные, теоретически обоснованные методы решения подобных задач пока имеют чрезмерно вы-

сокую вычислительную сложность, делающую невозможным их применение для решения широкого круга прикладных задач.

Однако одного наличия эффективных с точки зрения теории оптимизации методов недостаточно для решения сложных прикладных задач. На современном этапе для этого применяются программные продукты, которые реализуют различные многометодные схемы, включающие комбинации различных численных алгоритмов. Важным моментом, о котором нельзя забывать, является факт того, что алгоритм метода и его конкретная программная реализация, вообще говоря, являются разными объектами. Это сказано как с различным языком описания и, соответственно, уровнем применяемых абстракций, так и с ограничениями, накладываемыми аппаратной архитектурой, которая используется для исполнения созданного программного кода. В силу целого ряда технических факторов прекрасный с математической точки зрения метод может оказаться малоэффективным в рамках используемого вычислительного устройства. Важность этой проблемы хорошо известна и изучена многими ведущими специалистами, работающими в области численных методов. К настоящему времени разработано большое число различных коммерческих и открытых программных продуктов и библиотек, реализующих популярные алгоритмы локальной и глобальной оптимизации: Matlab, Maple, Mathematica, NAG, AMPL, CVX, NLopt и множество других. Но новые вызовы, которые бросают современные постановки конечномерной оптимизации большой размерности, а также бурное развитие новых архитектур вычислительной техники, заставляют искать новые, более эффективные способы и подходы решения таких задач. Многие из алгоритмов оптимизации, используемых в популярных математических пакетах, уже не могут быть признаны наилучшими на доступных вычислительных системах и не могут полностью использовать предоставляемые вычислительные мощности. Попытке решения этой проблемы и конструированию многометодных алгоритмов с характеристиками, адекватными современным вычислительным архитектурам, и посвящена диссертационная работа.

Объектом исследования являются алгоритмы и вычислительные технологии решения задач локальной и глобальной оптимизации большой размерности.

Целью работы является поиск эффективных сочетаний современных вычислительных архитектур и оптимизационных методов для решения актуальных задач максимально возможных размерностей. Вместо выбора «наилучшего» метода должен производиться выбор наиболее эффективной пары «метод + архитектура» – т.е. алгоритм и его программная реализация должны учитывать аппаратные особенности выбранной вычислительной платформы. Явный учёт специфики архитектуры позволяет добиться существенно лучших результатов, чем при использовании общих («универсальных») подходов.

Для достижения поставленной цели были решены следующие задачи:

1. Определение предельных размерностей, достижимых на современных вычислительных архитектурах.
2. Селекция известных методов, наиболее подходящих для создания параллельных реализаций.
3. Разработка новых оптимизационных методов и их модификаций.
4. Разработка вычислительных технологий оптимизаций для современных высокопроизводительных архитектур.
5. Разработка методик стандартизации программного кода градиентных алгоритмов оптимизации.
6. Проверка работоспособности и эффективности выполненных программных реализаций путём решения различных модельных и прикладных задач локальной и глобальной оптимизации.

Методы исследования. Для решения поставленных задач используются методы математического программирования, невыпуклой оптимизации, теории экстремальных задач, методики тестирования, методы проектирования программного обеспечения и технологии параллельного программирования.

Достоверность и обоснованность полученных в диссертационной работе результатов обеспечивается: тестированием на коллекциях общепризнанных модельных задач с известными оптимальными решениями, в том числе

аналитическими; контролем выполнения необходимых условий оптимальности для найденных решений; верификацией результатов расчётов с применением многометодных схем – решение каждой задачи найдено с использованием различных оптимизационных алгоритмов с последующим анализом «разброса» (полученных) результатов; сравнением результатов вычислительных экспериментов с расчётами, проведёнными другими авторами (исследователями); содержательными трактовками найденных решений экспертами и постановщиками задач.

Научная новизна результатов работы, выносимых на защиту:

1. Разработаны новые варианты покомпонентных методов оптимизации аддитивных функций, получены оценки скорости сходимости.
2. Предложены новые методики ускорения оптимизационных алгоритмов, явно учитывающих структурные особенности решаемых задач.
3. Разработана и верифицирована методика стандартизации программного кода для CPU- и GPU-реализаций градиентных алгоритмов.
4. На основе метода Б.Т. Поляка разработана новая вычислительная технология, позволившая за счёт учёта аппаратных характеристик современных вычислительных кластеров повысить размерность решаемых задач до 10^{11} переменных.
5. Разработан новый GPU-ориентированный алгоритм глобальной оптимизации («Forest»).
6. Получены решения новых прикладных задач из областей молекулярного моделирования, анализа интернет-сетей и транспортного моделирования.

Научная и практическая значимость. Результаты, представленные в диссертационной работе, носят как теоретический, так и практический характер. Предложены и исследованы, теоретически и экспериментально, новые варианты покомпонентных методов, а также новые методики ускорения их алгоритмов. Для ряда прикладных задач получены новые, не известные ранее, решения. Разработано специализированное программное обеспечение,

внедрённое в нескольких научно-исследовательских и производственных организациях. Результаты диссертационных исследований также могут быть использованы при решении широкого круга задач из областей анализа данных, конструирования новых материалов, проектирования транспортных систем.

Основные результаты работы, выносимые на защиту:

1. Новые варианты покомпонентных методов локальной оптимизации и оценки скорости их сходимости.
2. Новые методики ускорения оптимизационных алгоритмов, явно учитывающих структурные особенности решаемых задач.
3. Методика стандартизации программного кода градиентных алгоритмов, использующих многопроцессорные системы и графические ускорители.
4. Новый алгоритм глобальной оптимизации («Forest»), эффективно использующий мощности современных GPU.
5. Решения задач рекордных размерностей из областей молекулярного моделирования, анализа данных и транспортного моделирование.

Апробация работы. Результаты, представленные в диссертации, докладывались и обсуждались на ведущих международных и российских конференциях:

- 4-я международная конференция «Квазилинейные уравнения, обратные задачи и их приложения» (QIPA-2019), МФТИ, Долгопрудный, 3–5 декабря 2019 г.;
- 19-я Всероссийская конференция «Математические методы распознавания образов», Москва, 26–29 ноября 2019 г.;
- 62-я Всероссийская научная конференция МФТИ, Долгопрудный, 18–23 ноября 2019 г.;

- International Conference «Optimization and applications» (OPTIMA), 2011–2016, 2018 гг.;
- Международная школа-семинар «Нелинейный анализ и экстремальные задачи», 2014–2018 гг.;
- XVII Байкальская международная школа-семинар «Методы оптимизации и их приложения», с. Максимиха, Республика Бурятия, 31 июля – 6 августа 2017 г.;
- Международная конференция «Дискретная оптимизация и исследование операций», Владивосток, остров Русский, 19–23 сентября 2016 г. (соавтор доклада);
- Joint ORSC/EURO International Conference 2015 on Continuous Optimization, Шанхай, Китай, 10–12 мая 2015 г. (соавтор доклада);
- 27th European Conference on Operational Research, University of Strathclyde, Глазго, Великобритания, 12–15 июля 2015 г. (соавтор доклада);
- 17-я Всероссийская конференция «Математические методы распознавания образов», г. Светлогорск, Калининградская область, 25 сентября 2015 г. (соавтор пленарного доклада);
- International conference ITAS-2015, Олимпийская деревня, Сочи, 7–11 сентября 2015 г. (соавтор доклада);
- Конференция «Выпуклая оптимизация и “стохастические” приложения», ИППИ РАН, Москва, 23 декабря 2015 г.;
- Отчётная транспортная конференция, МЦНМО, Москва, 26 декабря 2015 г.;
- XX Байкальская Всероссийская конференция и Школа-семинар научной молодёжи, Иркутск - Байкал, 29 июня – 7 июля 2015 г. (соавтор пленарного доклада);

- Десятая международная азиатская школа-семинар «Проблемы оптимизации сложных систем», Кыргызская республика, 25 июля – 5 августа 2014 г. (соавтор доклада);
- Systems Dynamics and Control Processes, International Conference dedicated to the 90th Anniversary of Academician N.N. Krasovskii, Екатеринбург, 15–20 сентября 2014 г. (соавтор доклада);
- Шестая традиционная всероссийская молодёжная летняя школа «Управление, информация и оптимизация» (ИПУ РАН), дер. Григоричово, Московская область, 22–29 июня 2014 г.;
- The 4th International Conference on Optimization, Simulation and Control, Улан-Батор, Монголия, 1–4 июля 2013 г.;
- XV Байкальская международная школа-семинар «Методы оптимизации и их приложения», п. Листвянка, Иркутск–Байкал, 23–29 июня 2011 г.;
- 4-я, 6-я Международная научная конференция «Параллельные вычислительные технологии (ПаВТ)», 2010, 2012 гг.;
- 11-я, 13-я Всероссийская конференция молодых учёных по математическому моделированию и информационным технологиям, 2010, 2012 гг.

Результаты работы также докладывались и обсуждались на научных семинарах академика Ю.Г. Евтушенко (ФИЦ ИУ, Москва), академика Б.Н. Четверушкина (ФИЦ ИПМ РАН, Москва), академика К.В. Рудакова (ФИЦ ИУ, Москва), чл.-корр. А.В. Двуреченского (ИФП СО РАН, Новосибирск), проф. А.П. Афанасьева (ИППИ РАН, Москва), проф. М.Г. Дмитриева (ИСА РАН, Москва), семинарах ИММ УРО РАН (Екатеринбург) и ИВТ СО РАН (Новосибирск).

Результаты, представленные в диссертационной работе, использовались для выполнения работ по проекту Министерства образования и науки РФ «Разработка архитектуры, принципов построения и программного обеспечения суперкомпьютерной информационно-аналитической платформы при-

кладного моделирования, прогнозирования и экспертиз транспортных процессов и систем в виде сетевой компьютерной лаборатории» (2015–2016 гг., руководитель: академик Б.Н. Четверушкин, ИПМ РАН), проектам РФФИ 2-07-31118 мол_а (2012–2013 гг, руководитель: А.С. Аникин), 18-31-20005 (2018–2020 гг, руководитель: Гасников А.В. МФТИ), 18-29-03071 (2018–2020 гг, Нурминский Е.А. ДВФУ), 18-07-00587 (2018–2019 гг, Горнов А.Ю., ИДСТУ СО РАН), 17-07-00627 (2017–2019 гг, Зароднюк Т.С., ИДСТУ СО РАН), 16-07-00664 (2016 г, руководитель: А.Н. Андрианов, ИПМ РАН), 15-07-03827 (2015–2016 гг, руководитель: А.Ю. Горнов, ИДСТУ СО РАН), 13-01-00470 (2013–2015 гг, руководитель: А.Н. Андрианов, ИПМ РАН), 10-01-00595 (2010–2012 гг, руководитель: А.Н. Андрианов, ИПМ РАН), интеграционным проектам СО РАН №43 (2011 г, руководитель: чл.-корр. РАН А.В. Двуреченский, ИФП СО РАН) и №83 (2012 г, руководитель: чл.-корр. РАН А.В. Двуреченский, ИФП СО РАН), а также проекту «Traffic matrix estimation» (2015 г., договор между МФТИ и компанией Huawei Russia).

Личный вклад соискателя. Все исследования, изложенные в диссертационной работе, проведены лично соискателем в процессе научной деятельности. Из совместных публикаций в диссертацию включён лишь тот материал, который непосредственно принадлежит соискателю.

Публикации.

Основные результаты по теме диссертации изложены в 57 печатных изданиях, 13 из которых опубликованы в журналах из перечня Web of Science/Scopus/БАК ([1–13]) , 40 – в тезисах докладов.

Объем и структура работы. Диссертация изложена на 180 страницах и состоит из введения, четырёх глав и заключения. Работа содержит 37 рисунков, 18 таблиц и 121 наименование списка литературы.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность работы, формулируются цели и задачи исследования, основные защищаемые положения, а также приводятся данные о структуре и объёме диссертационной работы.

В первой главе диссертационной работы рассматриваются основные трудности, возникающие при решении современных задач конечномерной оп-

тимизации. Рассматриваются детали того, как именно указанные факторы влияют на процесс поиска решения задачи и возможные пути их преодоления. В качестве одного из способов предлагается использование параллельных вычислительных устройств, обеспечивающих как существенное ускорение процесса поиска решения, так и возможность решения задач, которые не могут быть обработаны на персональных современных вычислительных системах из-за аппаратных ограничений. Рассматриваются два основных способа организации оптимизационных алгоритмов — распараллеливание вычислений, связанных с задачей (вычисление значения функции, её градиента и т.п.), либо непосредственно распараллеливание алгоритма выбранного метода оптимизации. Для каждого из указанных способов описывается преимущественная область применимости, как с точки зрения количественных характеристик задачи (например, размерности), так и с точки зрения класса постановки — задача локальной либо глобальной оптимизации.

Вторая глава диссертационной работы посвящена исследованию современных параллельных вычислительных архитектур с точки зрения их применимости для решения рассматриваемого класса задач конечномерной оптимизации. Для каждой рассмотренной в главе аппаратной платформы выполнена оценка класса задач и алгоритмов оптимизации, которые целесообразно реализовывать с её использованием. При выполнении таких оценок учтены различные факторы: аппаратные возможности и ограничения, технологическая зрелость платформы, её распространённость на практике, уровень поддержки со стороны производителей аппаратного обеспечения, уровень поддержки со стороны программной составляющей (доступные технологии параллельного программирования, поддержка различных языков программирования, наличие развитых библиотек алгоритмов и т.п.), а также оценки перспектив дальнейшего развития.

Третья глава диссертационной работы посвящена программной реализации предложенных в работе методов и подходов. Реализация вычислительных модулей выполнена на языке C++, является кроссплатформенной и поддерживает использование наиболее популярных параллельных вычислительных платформ — систем с общей памятью, систем с распределённой памятью (кластеры) и графических ускорителей (GPU). Раздел 3.1 описывает выбран-

ные для реализации программные средства, включающие средства разработки, отладки, профилирования, а также набор внешних программных библиотек. Раздел 3.2 посвящён базовым структурам данных, реализованным в процессе разработки:

1. Плотные (dense) вектора — автором выполнена реализация для алгоритмов, использующих возможности как центрального процессора (CPU), так и графического ускорителя (GPU);
2. Разреженные (sparse) матрицы — автором выполнена собственная CPU–GPU реализация широко распространённого на практике формата хранения CSR (Compressed Sparse Row);
3. Бинарные вычислительные деревья — структуры, предназначенная для использования в рамках покомпонентных методов и позволяющая выполнять обновление значения некоторой древовидно-представимой функции вместо полного её вычисления в случае изменения отдельных компонент вектора переменных. Идея применения таких деревьев при решении больших задач оптимизации была предложена Ю.Е. Нестеровым. Автором диссертационной выполнена высокоэффективная программная реализация вычислительных деревьев, позволяющая производить требуемую процедуру обновления (пересчёта) значения требуемой функции. Реализация выполнена в обобщённом виде и может быть использована для построения деревьев, обрабатывающих входные данные различных типов. В качестве примера функции, которая может быть реализована с использованием рассматриваемого подхода, может быть выбрана некоторая норма (см. также рис. 1):

$$f(x) = \|x\|_p \rightarrow \begin{cases} \hat{f}^0(t_1, t_2) = (|t_1|^p + |t_2|^p), \\ \hat{f}^h(t_1, t_2) = (t_1 + t_2), 1 \leq h < k, \\ \hat{f}^k(t) = (t)^{1/p}, \\ f(x) = \hat{f}^k(v_1^k). \end{cases}$$

$$k = \lceil \log_2 n \rceil \geq \log_2 n.$$

$$v_j^{h+1} = \hat{f}^h(v_{2j-1}^h, v_{2j}^h).$$

v_1^4								
v_1^3								v_2^3
v_1^2				v_2^2				v_3^2
v_1^1		v_2^1		v_3^1		v_4^1		v_5^1
v_1^0	v_2^0	v_3^0	v_4^0	v_5^0	v_6^0	v_7^0	v_8^0	v_9^0

Рис. 1. Бинарное вычислительное дерево длины 9.

Применение таких структур позволяет радикально ускорить пересчёт подобных функций, что имеет огромное значение при решении задач с большими и сверхбольшими размерностями, что полностью подтверждается результатами вычислительных экспериментов. Автором диссертационной работы предложен и реализован механизм *отсечений*, позволивший на практике существенно улучшить поведение вычислительных деревьев за счёт ускорения процедуры обновления.

Раздел 3.3 посвящён описанию базовых алгоритмов, используемых при реализации методов оптимизации:

1. BLAS–процедуры — автором реализован набор базовых алгоритмов линейной алгебры, предназначенных для работы в рамках CPU–GPU алгоритмов.
2. Разностные схемы вычисления градиента.
3. Быстрый генератор дискретной случайной величины — реализация выполнена с использованием технологии бинарных вычислительных деревьев, что обеспечивает стоимость генерации отдельной дискретной величины порядка $O(\ln n)$.
4. Схемы прямого обновления сумм — автором предложен и реализован ряд схем, позволяющих производить обновление (пересчёт) значения функции, которая является суммой некоторых величин, за константное время (т.е. за $O(1)$). Независимость таких схем от размерности исходной задачи имеет огромное значение при работе со сверхбольшими задачами и позволяет радикально повысить быстродействие оптимизационных

алгоритмов. Результаты вычислительных экспериментов показали полную оправданность и работоспособность предложенного подхода.

В разделе 3.4 описывается реализация используемых методов одномерной минимизации — модифицированная версия комбинированного метода Брента, а также двух- и трёх-точечные варианты метода парабол, выполненные как в традиционном варианте, так и в покомпонентной версии, позволяющей использовать их в соответствующих оптимизационных методах.

В разделе 3.5 рассматривается параллельная реализация основных градиентных методов локальной оптимизации: различные варианты метода сопряжённых градиентов, метод Barzilai–Borwein и метод Б.Т. Поляка. Реализация всех алгоритмов выполнена с использованием предложенной автором диссертационной работы технологии, унифицирующей код алгоритмов, использующих вычислительные возможности CPU и GPU. Результаты проведённых вычислительных экспериментов продемонстрировали высокую эффективность предложенного подхода, а также высокий потенциал дальнейшего его развития и применения как для других алгоритмов, так и для других параллельных архитектур.

Раздел 3.6 диссертационной работы посвящён реализации современных покомпонентных методов локальной оптимизации. Основой идеологии таких методов является оптимизация на каждой итерации лишь 1-2 переменных и использование при этом процедур пересчёта значения функции вместо «традиционного» полного вычисления. Т.е. такие методы подразумевают выполнение большого числа «лёгких» итераций за счёт правильного учёта либо разреженности исходной постановки, либо разреженности решения. В диссертационной работе реализованы:

1. hNL1 (h Nesterov L1) — покомпонентная модификация прямого градиентного метода в 1-норме.
2. hRPDS — покомпонентный метод «диагонального поиска» hRPDS (h Randomized Parabolic Diagonal Search). Метод предложен автором диссертационной работы, являясь результатом развития идей, заложенных в основу метода hNL1, и предназначен для минимизации квадратичных

функций, но не требует при этом обязательного вычисления градиента:

$$f(x) \rightarrow \min_{x \in X}, \quad X = \left\{ x \in \mathbb{R}^n \mid \sum_{i=1}^n x_i = N \right\}.$$

Основной идеей представленного метода является «хождение вдоль диагоналей» (чем и обусловлено выбранное название), позволяющее производить поиск минимума оптимизируемой функции без нарушения имеющихся ограничений. Предложенный покомпонентный метод может быть расширен на более общий класс оптимизируемых функций путём реализации соответствующей процедуры алгоритма одномерного поиска. Возможно и применение более сложных процедур рандомизации, например с использованием информации о величине, на которую удалось уменьшить целевую функцию при выборе конкретных переменных — это можно сделать, например, с применением генератора дискретной случайной величины. Стоит также отметить, что достаточно простым образом в метод может быть добавлена параллелепипедных ограничений.

3. hFW (h Frank-Wolfe) — покомпонентная модификация метода условного градиента (Франк-Вульфа).
4. hRPCD (h Randomized Parabolic Coordinate Descent) — покомпонентная модификация метода покоординатного спуска. Предлагаемый автором работы метод предназначен для минимизации квадратичных задач безусловной оптимизации:

$$\begin{aligned} f(x) &= \|Ax - b\|_2^2 \rightarrow \min_{x \in \mathbb{R}^n}, \\ f(x) &= xAx - b \rightarrow \min_{x \in \mathbb{R}^n}. \end{aligned}$$

Для задач такого класса доказаны оценки скорости сходимости:

$$\phi^k - f^* = \frac{2n}{k+4} \cdot R_1^2(x^0),$$

где $\phi^k = E_{\xi^k} f(x^k)$, $\xi^k = \{i^0, i^1, \dots, i^k\}$ — некоторая случайная последовательность оптимизируемых переменных, выбранных методом за k итераций и $R_1(x^0) = \max_{x \in \mathbb{R}^n} \left\{ \max_{x^* \in X^*} \|x - x^*\|_0 : f(x) \leq f(x^0) \right\}$.

Полученный метод показал свою эффективность при решении ряда прикладных задач большой размерности, в том числе, задачи восстановления матрицы корреспонденций компьютерных сетей.

5. hRCHJ (h Randomized Coordinate Hooke–Jeeves) — покомпонентная модификация метода покоординатного спуска в варианте Хука–Дживса. Метод Хука–Дживса был предложен в 1961 году и задал вектор развития целому семейству т.н. методов «поиска по образцу» Оригинальная версия метода предполагала выполнение двух фаз алгоритма:

- (a) Исследующий поиск (покоординатный спуск).
- (b) Поиск по образцу.

В работе автором предлагается вариант покомпонентного метода, основанный на модифицированном варианте первой фазы (исследующем поиске). Предложенный покомпонентный метод, дополняет идею «сжатия» шага координатного спуска, предложенную в оригинальном работе Хука–Дживса, операцией его «расширения». Такой подход повышает гибкость метода, поскольку позволяет ему адаптировать своё поведение в зависимости от свойств целевой функции в конкретной области поиска. Также это позволяет методу сохранять свою эффективность в случае, например, задания заниженного значения начального шага h_0 .

Предложенный метод может быть успешно применён для решения, в первую очередь недифференцируемых, задач оптимизации большой размерности.

6. hARCD (h Accelerated Randomized Coordinate Descent) – предлагаемый автором работы ускоренный покомпонентный метод для сильно разреженных задач выпуклой оптимизации. К настоящему времени известны ускоренные варианты покомпонентных методов, но к сожалению, для некоторых классов гладких задач выпуклой оптимизации не очень понятно, как их применять, чтобы стоимость итерации учитывала раз-

реженность исходной задачи. Например, для постановок вида

$$\ln \left(\sum_{k=1}^m \exp (\langle a_k, x \rangle) \right) \rightarrow \min_{x \in \mathbb{R}^n}, \quad (1)$$

где матрица A , составленная из строк a_k , имеет в среднем в каждом столбце $s \ll m = O(n)$, отличных от нуля элементов, все известные ускоренные покомпонентные методы будут иметь стоимость итерации не меньше, чем $O(n)$, что может быть очень дорого при больших n . Автором предлагается варианта ускоренного покомпонентного метода, для которого амортизационная (т.е. средняя по всем итерациям) стоимость итерации будет $O(s)$.

Новый метод основан на комбинации известного неускоренный рандомизированного покомпонентного метода Ю.Е. Нестерова и техники ускорения «Каталист» (Catalyst). Для задачи (1) получены оценки общего числа итераций:

$$\tilde{O} \left(\sqrt{n \sum_{i=1}^n L_i R^2 / \varepsilon} \right), \quad (2)$$

при сложности итерации равной $O(s)$. Такая оценка (по числу итерации) хуже, чем у известных ускоренных покомпонентных методов, но сложность итерации при этом существенно зависит от свойств разреженности решаемой задачи. Предлагаемый в работе метод должен быть эффективен для сильно разреженных постановок, когда потери из-за возросшего числа итераций с запасом компенсируются их низкой стоимостью.

Раздел 3.7 посвящён рассмотрению реализованных методов поиска глобального экстремума:

1. MSBH — параллельная модификация хорошо зарекомендовавшего себя на практике метода Basin-Hopping. Метод относится к классу стохастических и реализует идею двухуровневого поиска — на верхнем производится случайная «стрельба» по всей доступной области поиска с последующими локальными спусками, а на нижнем — плотная «стрельба» со

спусками уже в небольшой области вокруг найденного ранее локального экстремума. Автором выполнена параллельная реализация метода MSBH, предназначенная для использования в рамках современных многопроцессорных систем (SMP). Для повышения эффективности поиска глобального экстремума решаемой задачи с использованием технологии OpenMP производится запуск параллельных вычислительных потоков, каждый из которых независимо выполняет как первый, так и второй уровень алгоритма. Предложенный вариант также позволяет без существенных затрат реализовать эффективную версию алгоритма для систем с распределённой памятью, что является важным фактором при решении сложных постановок глобальной оптимизации.

Предложенная реализация метода была успешно применена для решения многих задач глобальной оптимизации, включая поиск минимума потенциала Морса, которая многими ведущими специалистами в области оптимизации относится к классу весьма сложных постановок.

2. «Forest» — метод предложен автором диссертационной работы на основе опыта, полученного при численном решении различных задач глобальной оптимизации. При его разработке была осуществлена попытка создания алгоритма, который бы изначально был ориентирован на параллельную работу и пытался делать это максимально эффективно. Ещё одним важным стимулом к началу работ в этом направлении послужил успешный опыт применения вычислительных мощностей современных GPU при решении ряда задач других классов и желание распространить этот опыт и на задачи глобальной оптимизации.

Основной идеей предложенного метода является разбиение параллельно запущенных локальных спусков на сегменты фиксированного «размера» (см. рисунок 2) и выполнение анализа полученных промежуточных результатов для «отсеивания» неэффективных экземпляров при помощи выбранных (эвристических) критериев. Такими критериями могут выступать:

- Традиционные критерии останова, т.е. достижение точки, обладающей требуемыми характеристиками, например, нормой градиента, меньшей, чем выбранное пороговое значение.
- Успешность того или иного спуска по сравнению с остальными. Возможно применение разнообразных вариантов такой оценки, например, по найденному значению функции в сравнении его со средним на текущий момент результатом. При этом, очевидно, необходимо учитывать «возраст» отдельного экземпляра для того, чтобы не отсеивать спуски, которые только начали свою работу.
- Близость найденных решений. В случае, когда несколько спусков приходят в одну и ту же область поиска, дальнейшая их одновременная работа, очевидно, неоправданна. Поэтому для дальнейшего поиска оставляется только экземпляр, достигший наилучшего результата, а остальные начинают спуски «с нуля». Минусом такого критерия является его вычислительная сложность, порядка $O(k^2)$, где k — число параллельных спусков, поэтому его применение требует осторожности и оправдано не во всех ситуациях.

Предложенный метод «Forest», таким образом, напоминает контролируемый процесс роста леса (чем и обусловлен выбор названия) — слишком старые деревья вырубаются, чтобы они не «занимали место» и не расходовали биоресурсы, расчищаются также слишком плотно «заросшие» области и производится вырубка отдельных деревьев, которые явно относятся к бесперспективным. Общей целью всех таких мероприятий является оптимизация использования ограниченных ресурсов, в нашем случае — вычислительных.

Предложенный метод был использован для решения ряда задач глобальной оптимизации и неплохо показал себя на практике. Например, при поиске минимумов потенциала Морса с применением мощностей современных графических ускорителей он продемонстрировал результаты, значительно превосходящие таковые для метода MSBH, который уже успел зарекомендовать себя как надёжный инструмент для задач рассматриваемого класса.

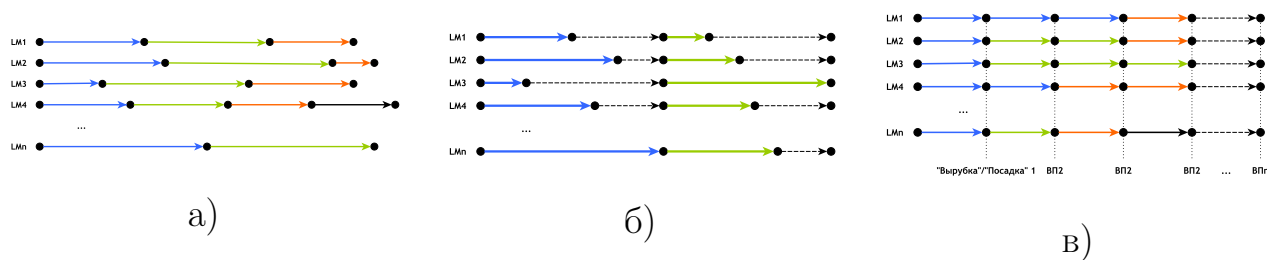


Рис. 2. а) — асинхронная работа локальных спусков; б) — синхронная работа локальных спусков; в) — синхронная работа локальных спусков в методе Forest.

Глава 4 диссертационной работы посвящена решению ряда прикладных и модельных задач конечномерной оптимизации большой размерности. С использованием предложенных в работе методов и подходов были решены следующие задачи:

1. Тестовые задачи локальной оптимизации класса Huge-Scale. В рамках проведения исследования с использованием параллельной версии модифицированного метода Б.Т. Поляка было получено решение ряда сепарабельных и квазисепарабельных задач выпуклой оптимизации с рекордными размерностями — порядка 10^{11} переменных. Вычислительные эксперименты производились на суперкомпьютерных системах ИПМ РАН и Межведомственного суперкомпьютерного центра РАН и показали хорошую масштабируемость предложенных подходов.
2. Задачи транспортного моделирования — поиск равновесного распределения транспортных потоков для модели Бэкмана. С использованием параллельных версий метода условного градиента было получено решение для ряда транспортных сетей с числом переменных порядка $\approx 10^5$. Предложены параллельные версии для систем с общей памятью и кластерных вычислительных систем, которые позволяют значительно (на 1-2 порядка) ускорить процесс нахождения равновесного распределения для сложных сетей, соответствующих уровню современных мегаполисов.
3. Задача восстановления матрицы корреспонденций компьютерных сетей, имеющая достаточно большое прикладное значение для многих

крупных компаний, занимающихся производством сетевого оборудования магистрального класса. Характерной особенностью актуальных вариантов задачи помимо их большой размерности является существенный объём оперативной памяти — крупные сети передачи данных требуют более 70 Гб ОЗУ для своего хранения. Дополнительной сложностью задачи является необходимость соответствия найденного решения ряду требований, определяемых т.н. «метриками качества». С использованием специализированной двухметодной вычислительной схемы, включающей применение градиентных и покомпонентных методов, автором было найдено решение ряда прикладных задач с числом переменных порядка 10^8 .

4. Задача ранжирования web-страниц (PageRank) — актуальный класс задач, эффективное решение которых имеет высокую прикладную ценность для многих современных интернет-компаний. Проведённые вычислительные эксперименты показали исключительную эффективность покомпонентных методов для рассматриваемого класса задач. В частности, для задачи web-Google с числом переменных порядка $9 \cdot 10^5$ решение может быть получено менее чем за 1 секунду.
5. Задача из области нанофизики — оптимизация атомно-молекулярного потенциала Китинга. С использованием предложенных в диссертационной работе параллельных методов и подходов были получены новые решения для ряда прикладных задач с числом переменных, превосходящим $2 \cdot 10^7$.
6. Задача из области математической химии — оптимизация атомно-молекулярного потенциала Морса. Рассматриваемая задача относится к классу задач поиска глобального экстремума и отличается очень высокой сложностью, обусловленной огромным числом локальных экстремумов. С использованием предложенных в работе параллельных версий методов MSBH и Forest получены новые решения для кластеров с числом атомов 241–250 (750 переменных). Найденные решения не публиковались ранее в работах других авторов, что позволяет считать их

новыми «рекордами», соответствующими общепризнанному в глобальной оптимизации принципу «Best-of-Known».

В заключении сформулированы основные результаты диссертационной работы:

1. Правильный выбор сочетания «метод + вычислительная архитектура» позволяет существенно ускорить поиск решения актуальных задач локальной и глобальной оптимизации.
2. Стандартные многопроцессорные системы (SMP) продолжают оставаться наиболее универсальным и доступным средством для численного решения широкого круга оптимизационных задач, включая некоторые постановки HugeScale-класса, с числом переменных до 10^9 .
3. Показано, что преодолев технологические сложности при создании GPU-реализаций оптимизационных алгоритмов, можно существенно повысить размерность решаемых задач.
4. Кластерные вычислительные системы становятся необходимыми при решении Huge-Scale задач с размерностями 10^{10} и более переменных, а также мультиэкстремальных задач глобальной оптимизации.
5. Для ряда задач класса LargeScale ($10^4 - 10^7$ переменных) конкурентоспособные результаты могут быть достигнуты вообще без применения технологий параллельного программирования.
6. Продемонстрирована принципиальная возможность решения задач локальной оптимизации с размерностями порядка $10^{11} - 10^{12}$ переменных на современных кластерных системах.

Литература

1. *Anikin A., Dorn Y., Nesterov Y.* Computational Methods for the Stable Dynamic Model // Optimization and Applications. Vol. 1145 / ed. by M. Jaćimović [et al.]. — Cham : Springer International Publishing, 2020. — P. 280–294. — (Communications in Computer and Information Science).
2. *Sorokovikov P., Gornov A., Anikin A.* Computational Technology for the Study of Atomic-Molecular Morse Clusters of Extremely Large Dimensions // IOP Conference Series: Materials Science and Engineering. — 2020. — Jan. — Vol. 734, 012092.
3. Methane Hydrate Emergence from Lake Baikal: Direct Observations, Modelling, and Hydrate Footprints in Seasonal Ice Cover / N. G. Granin [et al.] // Scientific Reports. — 2019. — Dec. 18. — Vol. 9, 19361.
4. A universal modification of the linear coupling method / S. Guminov [et al.] // Optimization Methods and Software. — 2019. — Vol. 34, no. 3. — P. 560–577.
5. A global optimization algorithm for solving a four-person game / S. Batabileg [et al.] // Optimization Letters. — 2019. — Vol. 13, no. 3. — P. 587–596.
6. Local Algorithms for Minimizing the Force Field for 3D Representation of Macromolecules / P. A. Yakovlev [et al.] // Computational Mathematics and Mathematical Physics. — 2019. — Dec. 1. — Vol. 59, no. 12. — P. 1994–2008.
7. *Kuz'menko V., Gornov A., Anikin A.* Estimation of Mathematical Models Accuracy for Calculation of LDL-Cholesterol Concentration // (Vth International Workshop "Critical Infrastructures: Contingency Management, Intelligent, Agent-Based, Cloud Computing and Cyber Security" (IWCI 2018)). — Atlantis Press, 2018. — P. 111–116.

8. Dual Approaches to the Minimization of Strongly Convex Functionals with a Simple Structure under Affine Constraints / A. S. Anikin [et al.] // Computational Mathematics and Mathematical Physics. — 2017. — Aug. 1. — Vol. 57, no. 8. — P. 1262–1276.
9. Numerical solution of huge-scale quasiseparable optimization problems / A. N. Andrianov [et al.] // Lobachevskii Journal of Mathematics. — 2017. — Vol. 38, no. 5. — P. 870–873.
10. Global Optimization Reduction of Generalized Malfatti's Problem / R. Enkhbat [et al.] // Numerical Algebra, Control & Optimization. — 2017. — Vol. 7, no. 2. — P. 211.
11. *Аникин А. С., Гасников А. В., Горнов А. Ю.* Рандомизация и разреженность в задачах Huge-Scale оптимизации на примере работы метода зеркального спуска // Труды МФТИ. — 2016. — т. 8, № 1. — с. 11–24.
12. *Аникин А. С., Гасников А. В., Горнов А. Ю.* О неускоренных эффективных методах решения разреженных задач квадратичной оптимизации // Труды МФТИ. — 2016. — т. 8, № 2. — с. 44–59.
13. Эффективные численные методы решения задачи PageRank для дважды разреженных матриц / А. С. Аникин [и др.] // Труды МФТИ. — 2015. — т. 7, № 4. — с. 74–94.