Введение

Это пособие написано прежде всего для студентов-математиков, начинающих изучать численные методы оптимизации и желающих впоследствии серьёзно погрузиться в данную область.

Пожалуй, основным численным методом современной оптимизации является *метод градиентного спуска*. Метод прекрасно изложен в замечательной книге Б. Т. Поляка [86], вышедшей в 1983 г. В некотором смысле этот метод порождает[1] большинство остальных численных методов оптимизации. Метод градиентного спуска активно используется в вычислительной математике не только для непосредственного решения задач оптимизации (минимизации), но и для задач, которые могут быть переписаны на языке оптимизации [16, 49, 86, 95, 292, 370] (решение нелинейных уравнений, поиск равновесий, обратные задачи и т. д.). Метод градиентного спуска можно использовать для задач оптимизации в бесконечномерных пространствах [507], например для численного решения задач оптимального управления [14, 48, 49, 86, 87, 292]. Но особенно большой интерес к градиентным методам в последние годы связан с тем, что градиентные спуски и их стохастические/рандомизированные варианты лежат в основе почти всех современных алгоритмов обучения, разрабатываемых в *анализе данных* [40, 82, 162, 177, 186, 202, 261, 336, 357, 403, 437, 493, 533, 547, 564, 608, 634].

♦ Всё это также хорошо можно проследить по трём основным конференциям по анализу данных: COLT, ICML, NIPS (NeurIPS), которые за последние 10–15 лет частично превратились в конференции, посвящённые использованию градиентных методов в решении задач *машинного обучения*. ♦

Не удивительно в этой связи, что подавляющее большинство современных курсов по численным методам оптимизации построено вокруг градиентных методов [13, 52, 76, 160, 164, 182, 186, 334, 336, 403, 437, 477, 495]. Данное пособие, подготовленное по материалам курса, прочитанного в ЛШСМ 2017, также построено по такому принципу.

[1] Собственно, данное пособие имеет одной из своих целей пояснить смысл этого предложения и слова «порождает» в данном контексте.

Однако принципиальное методическое отличие предложенного курса от остальных заключается в том, что в данном курсе предпринята попытка на примере только градиентного спуска продемонстрировать основной арсенал приёмов, с помощью которых разрабатываются новые численные методы и теоретически исследуется их скорость сходимости. Такое построение курса было обусловлено желанием в первую очередь донести основную идею того или иного приёма, не отягощая изложение техническими деталями. Градиентный спуск был выбран по нескольким причинам: во-первых, пожалуй, он самый простой, во-вторых, он лежит в основе большинства других методов, и если хорошо разобраться с тем или иным приёмом на примере градиентного спуска, то это можно использовать при перенесении на более сложный метод, лучше подходящий для решения конкретной задачи.

Курс начинается со стандартного изложения в § 1 того, что такое градиентный спуск. А именно, исходно сложная минимизируемая (целевая) функция заменяется в окрестности рассматриваемой точки касающимся её графика в этой точке параболоидом вращения, который по построению должен также мажорировать исходную функцию. Далее исходная задача минимизации заменяется задачей минимизации построенного параболоида. Последняя задача решается явно (осуществляется шаг градиентного спуска). Найденное решение задачи принимается за новую точку (положение метода), и процесс повторяется. В зависимости от того, какими свойствами обладала исходная функция (свойства гладкости, выпуклости), устанавливаются оценки на скорость сходимости описанной процедуры.

Начиная с § 2 изложение заметно усложняется, обрастая деталями. В § 2 рассматриваются задачи выпуклой оптимизации на множествах простой структуры (например, к таким множествам можно отнести неотрицательный ортант) в условиях небольших шумов неслучайной природы (см., например, [86, гл. 4]). Описанная выше процедура переносится на этот случай. Наличие шума играет ключевую роль в достижении одной из главных целей курса — построении *универсального градиентного спуска*. Этот метод сам настраивается на гладкость задачи и не требует параметров на входе.

В § 3 предлагается *концепция модели функции*, заключающаяся в том, что вместо параболоида вращения, аппроксимирующего (касающегося надграфика и мажорирующего) исходную выпуклую функцию в окрестности данной точки, можно использовать какие-то другие функции. Таким образом, например, можно дополнительно переносить «тяжесть» исходной постановки задачи на вспомогательные

подзадачи, надеясь, что это ускорит сходимость метода. Понятно, что такое ускорение будет достигнуто за счёт того, что каждая итерация станет дороже. Чтобы правильно выбрать по задаче модель функции, нужно иметь оценки того, насколько скорость сходимости внешней процедуры зависит от вида вспомогательных задач, точности их решения, и понимать, как сложность вспомогательных задач зависит от точности их решения. Всё это прорабатывается в данном параграфе при достаточно общих условиях.

В § 4 демонстрируется *прямодвойственная* природа обсуждаемых методов для выпуклых задач. Свойство прямодвойственности метода позволяет почти бесплатно получать решение задачи, двойственной к данной. Как правило, для большинства оптимизационных задач, приходящих из практики (экономика [17, 481, 482], транспорт [32, гл. 1, 3], проектирование механических конструкций [484] и даже анализ данных [32, гл. 5], [524]), двойственная задача несёт в себе дополнительную полезную информацию об изучаемом объекте (явлении), которую также хотелось бы получить в результате оптимизации. Другая не менее важная причина популярности прямодвойственных методов заключается в том, что, имея пару прямая–двойственная задача, можно выбирать, которую из них решать (какая проще). В частности, двойственные задачи являются задачами выпуклой оптимизации на множествах простой структуры. Если при решении выбранной задачи (прямой или двойственной) использовать прямодвойственный метод, то, решив её с некоторой точностью, гарантированно решим с такой же точностью и сопряжённую (двойственную) к ней задачу.

♦ Напомним, что при весьма общих условиях [182, гл. 5] двойственной задачей для двойственной к исходной выпуклой задаче будет исходная задача (теорема Фенхеля — Моро [66, п. 1.4, 2.2]). ♦

В § 5 строится *прямодвойственный универсальный градиентный спуск* для задачи выпуклой оптимизации на множестве простой структуры. Концепция универсального метода обобщает известное и популярное на практике правило выбора шага дроблением/удвоением [46, п. 6.3.2], см. также правила Армихо, Вулфа, Голдстейна [13, гл. 5], [56, п. 3.1.2], [59, п. 9.4], [76, п. 1.2.3], [86, гл. 3], [495, гл. 3], выбора шага градиентного спуска. Эта концепция подготавливалась около 30 лет (см., например, [71, 80]) и лишь весной 2013 г. была оформлена Ю. Е. Нестеровым сначала в виде препринта, а потом в виде статьи [492]. Статья вызвала большой интерес и сейчас активно цитируется в оптимизационном сообществе. Отличие универсального

подхода от *адаптивного* (к последнему можно отнести методы с выбором шага по отмеченным выше правилам типа Армихо) заключается в том, что настройка происходит не только на константу гладкости, но и на степень гладкости по шкале: негладкая → гёльдерова → гладкая функция. Универсальные прямодвойственные методы сейчас активно используются при поиске равновесий в больших транспортных сетях [7, 28, 32]. Большая популярность самонастраивающихся оптимизационных процедур в анализе данных, особенно в глубоком обучении[2] [40, 82, 134, 305] (в том числе использование нейросети для выбора величины шага в обучении другой нейросети), определённо указывает на то, что за адаптивными (самонастраивающимися), а по нашей терминологии «универсальными» методами будущее! Всё это, безусловно, также сильно сказалось на отборе материала и сделанных в пособии акцентах.

♦ Опыт использования терминов *прямодвойственный* и *универсальный* (см. [481, 492]) показывает, что оптимизационное сообщество в России принимает эти термины не однозначно. В частности, часто можно было слышать следующие замечания. «Представляется более естественным говорить про просто *двойственный метод* — см., например, метод Эрроу — Гурвица [86, п. 3, § 2, гл. 8], который имеет ещё более ярко выраженную прямодвойственную структуру, чем рассматриваемые в пособии, однако относится к классу *двойственных методов*. Словосочетание *универсальный метод* несколько вводит в заблуждение масштабами универсальности. Ведь в данном контексте речь идёт только об универсальном по гладкости методе, т. е. методе, который на вход не требует никакой информации о свойствах гладкости задачи (в том числе и константах, характеризующих гладкость). Однако, например, для сильно выпуклых задач такие методы требуют знания константы сильной выпуклости, и никакой самонастройки на эту константу по ходу работы (как в случае с константами, отвечающими за гладкость) уже не происходит». В целом, несмотря на эти замечания, было решено сохранить термины в неизменном виде, поскольку в англоязычной литературе они уже достаточно прочно

[2] Несмотря на огромную популярность этого направления и огромные усилия, затраченные на объяснение успешного практического опыта использования глубоких нейронных сетей в различных приложениях, важно подчеркнуть, что на данный момент, насколько нам известно, учёные по-прежнему достаточно далеки от возможности научно всё это объяснить, в том числе с точки зрения оптимизации. Более того, здесь имеются и вполне определённые отрицательные результаты [560].

успели закрепиться и их исправление может осложнить последующее изучение читателями современной литературы по данной тематике, которая в основном вся на английском языке. ♦

В приложении приводится краткий обзор современного состояния дел в активно развивающейся в последние годы области численных методов выпуклой оптимизации. Материал излагается в контексте результатов, приведённых в основном тексте пособия. Приложение написано в первую очередь для читателей, желающих продолжить изучение курса численных методов оптимизации. Надеемся, что приложение поможет читателям сориентироваться и укажет на некоторые новые направления и возможности.

Важную роль в тексте пособия играют замечания и упражнения, которые рекомендуется как минимум просматривать, а лучше прорешивать. В частности, таким образом (через замечания и упражнения) вводятся два основных приёма (сохраняющих оптимальность методов в смысле числа обращений к оракулу), позволяющих переходить от выпуклых задач к сильно выпуклым и обратно, соответственно *метод регуляризации* и *метод рестартов*. Имея метод, настроенный на сильно выпуклые задачи с помощью регуляризации функционала, можно привести любую задачу к сильно выпуклой и использовать имеющийся метод. Обратно, имея метод, настроенный на выпуклые задачи, можно использовать данный метод для решения сильно выпуклых задач, *рестартуя* (перезапуская) его каждый раз, когда расстояние до решения сокращается в два раза. В упражнениях также обсуждаются *ускоренный* градиентный (*быстрый*, *моментный*) спуск и *теория нижних оракульных оценок* сложности задач выпуклой оптимизации, построенная в конце 70-х годов XX века А. С. Немировским и Д. Б. Юдиным [74]. В замечании 3.3 описывается общий способ (*каталист*) ускорения неускоренных методов любого порядка.

В пособии имеется также несколько исторических замечаний и замечаний «второго плана», выделенных следующим образом:

♦ … ♦.

Изложение построено таким образом, что по ходу изучения материала должна появляться интуиция о возможности практически произвольным образом и в любом количестве сочетать различные описанные приёмы (конструкции, надстройки) друг с другом, получая таким образом всё более и более сложные методы, лучше подходящие под решаемую задачу. В этой связи, наверное, можно сказать, что в пособии описаны «структурные блоки», из которых строятся современ-

ные градиентные методы. Замечательно, что эти же структурные блоки используются и для ускоренных методов и их стохастических и рандомизированных вариантов, см. приложение, а также [16, 20, 21, 31, 32, 75, 76, 98, 120, 127, 164, 233, 271, 272, 352, 355, 375, 402, 403, 469, 487, 492].

Приведём здесь для удобства основные структурные блоки (приёмы) для методов первого порядка (градиентных методов) с указанием частей пособия, в которых они описаны. Эти блоки переносятся и на методы другого порядка, однако детали мы вынуждены здесь опустить. Ограничимся также для простоты только четырьмя бинарными признаками, характеризующими решаемую задачу оптимизации и используемый метод:

1) задача гладкая/негладкая;
2) задача сильно выпуклая / выпуклая (вырожденная задача выпуклой оптимизации);
3) при решении задачи доступен градиент функционала / стохастический градиент;
4) для решения задачи используется ускоренный/неускоренный метод.

Далее (см. также табл. 2 в приложении и комментарии к ней) будут описаны приёмы, которые в совокупности позволяют по (оптимальному) алгоритму, отвечающему конкретному набору этих четырёх признаков, строить (оптимальный) алгоритм, отвечающий любому из пятнадцати оставшихся наборов этих признаков. Впрочем, необходимости строить по ускоренным методам неускоренные на практике не возникает, поэтому соответствующее описание далее опущено.

1. Негладкая задача может рассматриваться как гладкая за счёт искусственного введения неточности в параболическую модель аппроксимации оптимизируемой функции и адаптивной стратегии выбора кривизны параболической модели, см. § 5.
2. Любую выпуклую задачу можно сделать сильно выпуклой с помощью *регуляризации* (см. замечание 4.1), а любой алгоритм, настроенный на решение выпуклой задачи, можно использовать для решения сильно выпуклой задачи за счёт *рестартов*, см. упражнение 2.3 и конец § 5, а также приложение.
3. Стохастического оракула, выдающего градиент, можно свести к неточному (с малым шумом), но уже детерминированному оракулу с помощью *минибатчинга*, см. начало приложения. Идея приёма: возвращение вместо стохастического градиента оптимизируемой функции в рассматриваемой точке среднего арифметического независимых реализаций стохастических градиентов в этой же точке.

4. С использованием конструкции *каталист* (см. замечание 3.3), в основе которой лежит проксимальный ускоренный градиентный метод, можно ускорять произвольные неускоренные методы, предназначенные для решения задач гладкой сильно выпуклой оптимизации. При этом получаются ускоренные методы, сходящиеся согласно нижним оценкам с точностью до логарифмических множителей. Таким образом, в отличие от конструкций, описанных выше, в данной конструкции согласно теоретическим оценкам всё же приходится «заплатить» логарифмический множитель за «общность».

Отметим также, что все описанные выше конструкции могут быть рассмотрены в такой общности, как в § 4, 5, т. е. с более общей моделью и в прямодвойственном контексте.

Для более комфортного изучения материала пособия рекомендуется предварительно познакомиться с основами выпуклого анализа, например, в объёме одной из книг [66, 182, 527] и основами (вычислительной) линейной алгебры [95, 628].

Список литературы к пособию включает более 600 источников (при том, что мы далеко не всегда ссылались на первоисточники, в ряде случаев предпочитая более современные статьи и обзоры), поэтому вряд ли можно рассчитывать, что даже хорошо мотивированный читатель сможет ознакомиться с большей его частью. В этой связи для удобства выделим из этого списка учебники, изучение которых вместе с данным пособием можно рекомендовать в первую очередь.

I. *Boyd S.*, *Vandenberghe L.* Convex optimization. Cambridge University Press, 2004.

II. *Nocedal J.*, *Wright S.* Numerical optimization. Springer, 2006.

III. *Поляк Б. Т.* Введение в оптимизацию. М.: URSS, 2014. 392 с.

IV. *Bubeck S.* Convex optimization: algorithms and complexity // Foundations and Trends in Machine Learning. 2015. Vol. 8, № 3–4. P. 231–357.

Стэнфордский учебник [I] является наглядным и одновременно строгим введением в выпуклую оптимизацию (теорию двойственности, *принцип множителей Лагранжа* как следствие теоремы об отделимости гиперплоскостью граничной точки выпуклого множества от этого множества [66, п. 2.1], теоремы о дифференцировании функции максимума и т. п.), основы которой активно используются в настоящем пособии. Учебники [II, III] представляют собой достаточно подробное и хорошо проработанное описание основ численных ме-

тодов оптимизации (выпуклой и не выпуклой). Во многом на базе именно этих двух учебников происходит обучение студентов основам численных методов оптимизации в большинстве продвинутых учебных заведений по всему миру. Собранные в этих учебниках материалы отражают развитие данной области в основном в 60–80-е годы XX века. Более современные тенденции, связанные с развитием методов внутренней точки, ускорением методов и различными рандомизациями градиентных методов, отражены в учебнике Принстонского университета [IV]. Этот учебник можно рекомендовать в качестве основного источника для последующего изучения.

Во вторую очередь (для более строгого изучения предмета) можно рекомендовать следующие учебники.

a) *Ben-Tal A.*, *Nemirovski A.* Lectures on modern convex optimization analysis, algorithms, and engineering applications. Philadelphia: SIAM, 2019.
b) *Nesterov Yu.* Lectures on convex optimization. Springer, 2018.
c) *Lan G.* First-order and stochastic optimization methods for Machine Learning. Springer, 2020.

♦ С. Бойд [181] является сейчас одним из самых цитируемых и активно публикующихся учёных в области численных методов оптимизации. С. Бойд имеет инженерное образование и большое внимание в своих исследованиях уделяет практической составляющей, изящно сочетая её с фундаментальной. Оптимизационное сообщество практически едино во мнении, что работы С. Бойда (речь идёт прежде всего о его книгах и документациях к разработанным под его руководством пакетам типа CVX [621]) являются хорошим образцом ясности изложения. Курс [I] является, пожалуй, самым известным (востребованным) в последнее десятилетие курсом по выпуклой оптимизации. ♦

Отметим, что настоящее пособие довольно сильно отличается и по отбору материала, и по форме изложения от подавляющего большинства известных нам учебников по оптимизации, в том числе и от выделенных четырёх. Достаточно сказать, что в пособие не были включены ставшие уже классическими разделы про задачи линейного и полуопределённого программирования. Приведём здесь ссылки на то, как эти материалы в 2018/2019 учебном году преподавал А. С. Немировский студентам и аспирантам университета Джорджия в Атланте [461, 464]. Отметим также некоторые недавние достижения в этих областях [419, 420]. С другой стороны, почти половина из материалов пособия, по-видимому, впервые излагается (осмысляется) в учебном контексте.

В пособии имеется большое число ссылок на современную иностранную литературу. После распада СССР «оптимизационный крен» сильно сместился на Запад. Однако мы считаем важным подчеркнуть определяющую роль российских учёных и научных школ [512] в создании того фундамента, на котором сейчас стоит молодая (чуть больше 60 лет), но бурно развивающаяся область знаний: «численные методы оптимизации». На Западе даже есть такая вполне серьёзная шутка: «Если ты придумал новый численный метод оптимизации, не торопись радоваться, наверняка его уже знал какой-нибудь русский ещё в 60-е годы прошлого века и опубликовал, конечно, на русском языке». В частности, многое из того, что включено в данное пособие, было придумано нашими соотечественниками.

В 2004–2005 гг. автор, будучи студентом факультета управления и прикладной математики (ФУПМ) МФТИ, на базовой кафедре в ВЦ РАН слушал курс профессора В. Г. Жадана [52] по дополнительным главам численных методов оптимизации, оказавший заметное влияние на последующий интерес к этой области. В целом стоит отметить большое влияние школы акад. Н. Н. Моисеева на формирование как базового, так и дополнительного цикла оптимизационных дисциплин на ФУПМ [10, 11, 48, 49, 52, 68, 69]. Современный учебный план студентов ФУПМ состоит из сочетания отмеченного опыта школы Н. Н. Моисеева и опыта коллег с ВМиК МГУ [13, 14, 56, 59, 92]. В данном пособии предпринята попытка посмотреть на этот учебный план, формировавшийся в течение полувека, сквозь призму современных достижений в области численных методов выпуклой оптимизации [76, 164, 186] и новых приложений [32, 40, 608]. Отметим также практикумы [622, 626] к упомянутому циклу лекций для студентов ФУПМ.

Автор также постарался учесть и обыграть в пособии наработки, которыми с ним любезно делились на всевозможных конференциях и семинарах представители различных научных школ: В. П. Булатова (Иркутск), В. Ф. Демьянова (Санкт-Петербург), И. И. Ерёмина (Екатеринбург), Л. В. Канторовича (Санкт-Петербург, Новосибирск, Москва), М. М. Лаврентьева (Новосибирск), А. А. Милютина (Москва), В. А. Скокова (Москва), А. Н. Тихонова (Москва), Я. З. Цыпкина (Москва), Н. З. Шора (Киев), а особенно школ Ю. Г. Евтушенко (ВЦ РАН), Б. Т. Поляка (ИПУ РАН) и В. М. Тихомирова (мехмат МГУ). Вот уже более 10 лет автор имеет возможность обсуждать различные связанные с оптимизацией вопросы с Е. А. Нурминским, В. Ю. Протасовым, С. П. Тарасовым, С. В. Чукановым, А. А. Шананиным и А. Б. Юдицким.

Серьёзное влияние на автора оказало регулярное общение с 2011 г. с Б. Т. Поляком, А. С. Немировским и особенно с Ю. Е. Нестеровым. В большой части данный курс (пособие) был построен на расшифровке этих бесед. Автор очень благодарен трём оракулам за это.

Хотелось бы отметить важное влияние, которое оказала на данный текст совместная научная работа, выполняемая с А. Ю. Горновым, П. Е. Двуреченским, Ф. С. Стонякиным.

Автор также выражает благодарность своему коллеге по кафедре математических основ управления МФТИ доценту А. Г. Бирюкову за внимательное прочтение данной рукописи и предложенные исправления, а также Ф. Баху, Е. А. Воронцовой, К. В. Воронцову, А. И. Голикову, Н. В. Дойкову, Ю. В. Дорну, С. Э. Парсегову, А. О. Родоманову, Ф. Н. Рыбакову, Г. Скутари, Н. Сребро, А. Тейлору, Ц. Урибе, Р. Хильдебранду, А. В. Чернову за ряд ценных замечаний. На ряд неточностей автору было указано учениками: Артёмом Агафоновым, Мохаммадом Алкуса, Александром Безносиковым, Эдуардом Горбуновым, Сергеем Гуминовым, Дмитрием Камзоловым, Василием Новицким, Петром Остроуховым, Дмитрием Пасечнюком, Александром Рогозиным, Антоном Рябцевым, Абдурахмоном Садиевым, Даниилом Селихановичем, Александром Титовым, Даниилом Тяпкиным, Александром Тюриным, Ильнурой Усмановой.

Особую благодарность за постоянную поддержку хотелось бы выразить своей жене Даше Двинских.

Ответственность за все возможные ошибки лежит всецело на авторе. В случае обнаружения неточностей просьба присылать информацию на адрес электронной почты <gasnikov.av@mipt.ru>.