А. В. Гасников

Современные численные методы оптимизации

Метод универсального градиентного спуска

Учебное пособие

Москва
Издательство МЦНМО
2020

УДК 519.86(075)
ББК 22.17я73
Г22

Гасников А. В.

Г22 Современные численные методы оптимизации. Метод универсального градиентного спуска: учебное пособие. — М.: МЦНМО, 2020. — 272 с.

ISBN 978-5-4439-0000-1

В пособии рассматривается классический градиентный спуск. Однако изложение ведется на продвинутом уровне. Пособие отличается довольно полным обзором современного состояния методов типа градиентного спуска.

В книге делается акцент не на изложение методов, а на способы получения из старых методов новых с помощью небольшого числа общих приемов.

Пособие используется для преподавания численных методов оптимизации студентам 3 курса школы ПМИ МФТИ и 3 курса ФКН ВШЭ.

ББК 22.17я73

Рецензенты:

Директор Вычислительного центра им. А. А. Дородницына ФИЦ ИУ РАН академик РАН *Ю. Г. Евтушенко*

Директор Института вычислительной математики РАН, заведующий кафедрой вычислительных технологий и моделирования ВМиК МГУ академик РАН *Е. Е. Тыртышников*

ISBN 978-5-4439-0000-1

Оглавление

Предисловие Б. Т. Поляка

Трудно не согласиться с автором, когда он начинает книгу словами: «Пожалуй, основным численным методом современной оптимизации является *метод градиентного спуска*». Сейчас это кажется естественным и даже очевидным утверждением. Однако я хорошо помню времена, когда учебники по оптимизации начинались с описания симплекс-метода линейного программирования, а градиентному методу находилось место где-то на задворках. В то же время, несмотря на понимание возрастающей роли градиентного метода в современных приложениях, до сих пор не было ни одной монографии, специально посвящённой этому методу, демонстрации его богатства и широты возможностей.

Книга А. В. Гасникова закрывает этот зазор в научной литературе. Она написана так, что при желании её основные идеи и алгоритмы может понять новичок. В то же время она содержит массу тонкого и самого современного материала для продвинутого специалиста. В частности, в приложении такой читатель найдёт детальнейший обзор новейших численных методов оптимизации, обобщающих стандартный градиентный спуск, и их программных реализаций. Очень важно, что рассматриваются самые разнообразные классы задач (безусловная оптимизация и задачи с ограничениями, выпуклые и сильно выпуклые ситуации, детерминированные и стохастические модели и методы, гладкие и негладкие функции). По существу, несмотря на подзаголовок «метод универсального градиентного спуска», монография перерастает в универсальный учебник по методам оптимизации. Опытные лекторы смогут модифицировать стандартные курсы оптимизации, положив в их основу настоящую книгу.

Радостно сознавать, что в нашей стране появилось новое поколение специалистов по оптимизации, успешно продолжающих славные научные традиции.

Б. Т. Поляк

Предисловие

Данное пособие написано по материалам лекций, прочитанных автором в летней школе «Современная математика» в Ратмино (г. Дубна) в июле 2017 г.

Идея курса состояла в том, чтобы, с одной стороны, рассказать об основных приёмах, с помощью которых порождается многообразие современных численных методов выпуклой оптимизации первого порядка (рестарты, регуляризация, переход к двойственной задаче, адаптивная настройка на гладкость задачи, минибатчинг, каталист, слайдинг и т. д.). С другой стороны, хотелось провести все рассуждения на строгом математическом языке (с полным обоснованием). Поэтому для наглядности было решено ограничиться изучением только градиентного спуска и его окрестностей.

По данному пособию было прочитано несколько курсов лекций, которые снимались на видео: в осеннем семестре 2018/2019 в КМЦ АГУ [641], в весеннем семестре 2018/2019, 2019/2020 в школе ПМИ МФТИ, ФКН ВШЭ, а также в магистратуре MADE Mail.ru [640].

Опыт использования пособия при чтении лекций студентам школы ПМИ МФТИ показывает, что в пособие желательно ещё добавить подробно разобранные примеры решения практических задач оптимизации. Эту проблему планируется в перспективе решить за счёт издания другого пособия. Однако отметим, что даже в текущем варианте пособия можно найти примеры вполне реальных (практических) задач, с которыми мы сталкивались в разное время. Например, в конце § 5 и в примере из приложения рассматривается задача композитной оптимизации, возникшая при решении задачи восстановления матрицы корреспонденций в большой компьютерной сети по замерам потоков на линках (рёбрах). Эта обратная задача (обратные задачи являются естественным источником постановок задач в обычной оптимизации [370]) была поставлена компанией Хуавей в 2015 г.

Работа по подготовке пособия в § 1 и приложении была выполнена при поддержке Министерства науки и высшего образования Российской Федерации (госзадание № 075-00337-20-03, номер проекта 0714-2020-0005), в § 2 была поддержана грантом РФФИ 18-29-03071 мк, в § 3, 5 — грантом РФФИ 18-31-20005 мол_а_вед, в § 4 — грантом РФФИ 19-31-51001 Научное наставничество.

Обозначения

- $\mathbb{R}^n$ — n-мерное вещественное (векторное) пространство.
- $\mathbb{R}^n_+ = \{x \geqslant 0 \colon x \in \mathbb{R}^n\}$ — неотрицательный ортант пространства $\mathbb{R}^n$.
- const — числовая константа, значение которой зависит от контекста.
- $\dim x$ — размерность вектора x, в частности, $\dim x = n$, если $x \in \mathbb{R}^n$.
- $[x]_i$, x_i — i-я компонента вектора x.
- $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ — p-норма вектора $x = \{x_i\}_{i=1}^n \in \mathbb{R}^n$, $p \geqslant 1$.
- $\langle x, y\rangle = \sum_{i=1}^n x_i y_i$ — скалярное произведение векторов $x, y \in \mathbb{R}^n$. Отметим, что при определении графа используется похожее обозначение $G = \langle V, E\rangle$, имеющее другой смысл.
- $\|y\|_* = \sup_{\|x\| \leqslant 1} \langle x, y\rangle$ — сопряжённая норма к норме $\|\ \|$. В частности, для p-нормы сопряжённой будет q-норма, где $1/p + 1/q = 1$.
- $\lceil a \rceil = \max\{1, a\}$.
- $A \Rightarrow B$ — из утверждения (формулы) A следует утверждение (формула) B.
- $A \Leftrightarrow B$ — утверждение (формула) A эквивалентно (равносильно) утверждению (формуле) B, т. е. $A \Rightarrow B$ и $B \Rightarrow A$.
- $x \ll y$ — число y много больше числа x.
- $x \simeq y$, $x \approx y$ — число x приближённо равно числу y.
- $x \sim y$ — значение выражения x пропорционально значению выражения y, например, $V = HR \Rightarrow V \sim R$ или $T = 2\pi\sqrt{l/g} \Rightarrow T \sim \sqrt{l/g}$.
- $x := y$ — x присваивается y (пришло из программирования), например, $x := x + 1$.
- $\{z^1, \ldots, z^m\}$ — линейное пространство (подпространство в $\mathbb{R}^n$), «натянутое» на векторы $z^1, \ldots, z^m \in \mathbb{R}^n$, т. е. любой элемент такого пространства можно представить в виде

$$\alpha_1 z^1 + \ldots + \alpha_m z^m, \quad \alpha_1, \ldots, \alpha_m \in \mathbb{R}.$$

- A^T — матрица, транспонированная к матрице $A = \|A_{ij}\|_{i,j=1}^n$, т. е. $A^T = \|A_{ji}\|_{i,j=1}^n$.
- $\operatorname{tr}(A)$ — след квадратной матрицы A, т. е. сумма всех её диагональных элементов.

- $\operatorname{rank} A$ — ранг матрицы A, т. е. максимальное число линейно независимых столбцов (или строк).
- $\operatorname{nnz}(A)$ — число ненулевых элементов в матрице A.
- I — единичная матрица, т. е. $I = \|I_{ij}\|_{i,j=1}^{n}$, где $I_{ij} = 0$, если $i \neq j$, и $I_{ij} = 1$ иначе.
- A^{-1} — матрица, обратная к квадратной матрице A, т. е. $A^{-1}A = AA^{-1} = I$.
- $(\operatorname{Ker} A)^{\perp}$ — ортогональное дополнение подпространства, натянутого на собственные векторы матрицы A, отвечающие нулевому собственному значению.
- $\Im A$ — образ оператора A, т. е. множество таких y, для которых существует такой x, что $y = Ax$.
- $\sqrt{A}$ — квадратный корень из симметричной неотрицательно определённой матрицы A.

♦ Для каждой неотрицательно определённой симметричной матрицы существует такой ортонормированный базис, в котором действие этой матрицы можно понимать как соответствующее растяжение/сжатие/проектирование (задаваемое собственными значениями λ_i) вдоль ортов. Тогда действие матрицы $\sqrt{A}$ можно понимать как растяжение/сжатие/проектирование (задаваемое собственными значениями $\sqrt{\lambda_i}$) вдоль тех же самых ортов. ♦

- A^j — j-й столбец матрицы A; A_i — i-я строка матрицы A.
- $A \succ 0$ — симметричная матрица A ($A = A^T$) неотрицательно определена, т. е. $\langle x, Ax\rangle \geqslant 0 \ \forall\, x \in \mathbb{R}^n$.
- $A \succ B$ означает, что $A - B \succ 0$.
- $1_n = \underbrace{(1, \ldots, 1)}_{n}^{T}$ — вектор из единиц.
- $e_i = (\underbrace{0, \ldots, 0, 1}_{i}, 0, \ldots, 0)^T$ — i-орт.
- $S_n(1) = \left\{x \in \mathbb{R}^n_+ \colon \sum_{i=1}^n x_i = 1\right\}$ — единичный симплекс пространства $\mathbb{R}^n$.
- $B_{R,Q}(y) = \{x \in Q \colon \|x - y\|_2 \leqslant R\}$ — пересечение евклидова шара радиуса R с центром в точке y и множества Q.
- $\tilde{x} = \arg\min_{x \in P} F(x)$ означает, что $F(\tilde{x}) < F(x)$ для всех $x \in P \backslash \tilde{x}$.
- $\tilde{x} \in \operatorname{Arg}\min_{x \in P} F(x)$ означает, что $F(\tilde{x}) \leqslant F(x)$ для всех $x \in P$.

- $\pi_Q(x) = \arg\min_{y\in Q} \|x-y\|_2^2$ — евклидова проекция точки (вектора) x на замкнутое выпуклое множество Q.
- $\log a$ — логарифм положительного числа a по основанию, зависящему от контекста (в частности, $\ln = \log_e$). Если основание логарифма не указано, значит, основание зависит от контекста и это хочется подчеркнуть (см. приложение).
- $E_\xi[F(x,\xi)]$ — математическое ожидание по случайной величине (вектору) ξ от измеримой по ξ (вектор)-функции $F(x,\xi)$. Здесь x следует понимать как параметр.
- $E_\xi[f(\xi,\eta)\,|\,\eta]$ — условное математическое ожидание по случайной величине (вектору) ξ при «замороженной» случайной величине η от измеримой по ξ и η функции $f(\xi,\eta)$. Условное математическое ожидание является случайной величиной, зависящей от η.
- $\lambda_{\max}(A) = \max\{\lambda\colon \exists\, x\neq 0\colon Ax=\lambda x\}$,
  $\lambda_{\min}(A) = \min\{\lambda\colon \exists\, x\neq 0\colon Ax=\lambda x\}$.
- $\sigma_{\max}(A) = \lambda_{\max}(A^TA) = \lambda_{\max}(AA^T) = \max\{\lambda\colon \exists\, x\neq 0\colon AA^Tx=\lambda x\}$.
- $\widetilde{\sigma}_{\min}(A) = \min\{\lambda>0\colon \exists\, x\neq 0\colon AA^Tx=\lambda x\}$.
- $\nabla f(x) = \left(\dfrac{\partial f(x)}{\partial x_1},\ldots,\dfrac{\partial f(x)}{\partial x_n}\right)^T$ — градиент гладкой функции $f(x)$; аналогично определяется
$$\nabla_x f(x,y) = \left(\frac{\partial f(x,y)}{\partial x_1},\ldots,\frac{\partial f(x,y)}{\partial x_n}\right)^T.$$
- $\partial f(x)$ — субдифференциал выпуклой функции $f(x)$. По определению $g\in\partial f(x)$ (g — субградиент функции f в точке x) тогда и только тогда, когда для всех y имеет место неравенство
$$f(y)\geqslant f(x)+\langle g, y-x\rangle.$$
- $\nabla h(y) = \left\|\dfrac{\partial h_i(y)}{\partial y_j}\right\|_{i,j=1}^{n,m}$ — матрица Якоби гладкого отображения $h\colon \mathbb{R}^m\to\mathbb{R}^n$.
- $\nabla^2 f(x) = \left\{\dfrac{\partial \nabla f(x)}{\partial x_j}\right\}_{j=1}^n = \left\|\dfrac{\partial^2 f(x)}{\partial x_i\partial x_j}\right\|_{i,j=1}^n$ — матрица Гессе дважды дифференцируемой функции $f(x)$. Аналогично можно определить
$$\nabla^{r+1} f(x) = \left\{\frac{\partial \nabla^r f(x)}{\partial x_j}\right\}_{j=1}^n.$$
- $\nabla^{r+1} f(x)[u] = \sum_{j=1}^n \dfrac{\partial \nabla^r f(x)}{\partial x_j}\cdot u_j$ — тензор ранга r ($u\in\mathbb{R}^n$).
- $A(\text{параметры}) = O(B(\text{параметры}))$ означает, что существует такая абсолютная числовая константа C, не зависящая ни от каких па-

раметров, что

$$A(\text{параметры}) \leqslant C \cdot B(\text{параметры}).$$

- $A(\text{параметры}) = \widetilde{\mathrm{O}}(B(\text{параметры}))$ означает, что существует такой множитель $\widetilde{C}$, зависящий от параметров не сильнее, чем логарифмическим образом, что

$$A(\text{параметры}) \leqslant \widetilde{C} \cdot B(\text{параметры}).$$

Например,

$$\mathrm{O}\left(\sqrt{\frac{L}{\mu}} \ln \frac{\mu R^2}{\varepsilon}\right) = \widetilde{\mathrm{O}}\left(\sqrt{\frac{L}{\mu}}\right).$$

- $A \overset{\text{def}}{=} B$ означает, что $A = B$ и это равенство определяет либо A, либо B.

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>.

§ 1

Градиентный спуск

Рассмотрим задачу

$$f(x) \to \min_{x \in \mathbb{R}^n}. \tag{1.1}$$

Далее в этом параграфе приведены классические способы получения/понимания одного из основных инструментов современной вычислительной математики — *метода градиентного спуска*, восходящего к работам О. Коши [201], Л. В. Канторовича [58], Б. Т. Поляка [87]. Более современное изложение, в котором прорабатываются различные тонкие вопросы, начнётся со следующего параграфа.

♦ Стоит особо отметить большой (наверное, можно даже сказать, решающий) вклад, который внёс Б. Т. Поляк в 60-е годы XX века в развитие градиентных методов. Многие из современных методов и подходов, активно использующихся для решения задач оптимизации больших размеров, восходят к работам Бориса Теодоровича: усреднение Поляка [40, п. 8.7.3]; субградиентный метод Поляка [491]; *метод тяжёлого шарика* (импульсный метод), породивший впоследствии целую линейку ускоренных градиентных методов, в частности очень популярный в последние годы (быстрый, ускоренный, моментный) *градиентный метод Нестерова* (см. указание к упражнению 1.3). Собственно, знакомство с градиентными методами далее в пособии (особенно в §1, 2) осуществляется во многом под влиянием отмеченного цикла работ Б. Т. Поляка [86]. ♦

Рассмотрим систему обыкновенных дифференциальных уравнений [201]

$$\frac{dx}{dt} = -\nabla f(x). \tag{1.2}$$

Покажем, что значения функции $W(x) = f(x)$ убывают на траекториях динамической системы (1.2), т. е. $W(x)$ является *функцией Ляпунова* системы (1.2). Действительно,

$$\frac{dW(x(t))}{dt} = \left\langle \nabla f(x(t)), \frac{dx(t)}{dt} \right\rangle = \left\langle \nabla f(x(t)), -\nabla f(x(t)) \right\rangle = -\left\| \nabla f(x(t)) \right\|_2^2 \leqslant 0,$$

$$\frac{dW(x)}{dt} = 0 \iff \nabla f(x) = 0.$$

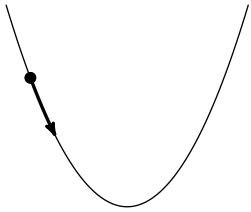

Рис. 1

Отсюда можно сделать вывод, что любая траектория такой системы должна сходиться к *стационарной точке*[1] функции $f(x)$, вообще говоря, зависящей от точки старта (на рис. 1 рассмотрен случай выпуклой функции). Аналогичного свойства можно ожидать и от дискретизованной по схеме Эйлера версии динамики (1.2)

$$x^{k+1} = x^k - h\nabla f(x^k) \tag{1.3}$$

в случае достаточно малого шага h [86, гл. 2], см. также [5, 13, 48, 74, 151, 240, 325, 455, 542, 572, 603]. Метод (1.3) обычно называют *методом градиентным спуском* или просто *градиентным спуском* [86], а приведённый здесь способ получения оценки скорости сходимости метода относят ко *второму методу Ляпунова* [86, § 2, гл. 2].

Чтобы количественно оценить скорость сходимости и получить условие на выбор шага, сделаем следующее предположение о *липшицевости градиента* в 2-норме [86, гл. 1]: для любых x и y имеет место неравенство

$$\|\nabla f(y) - \nabla f(x)\|_2 \leqslant L\|y - x\|_2. \tag{1.4}$$

Из этого неравенства следует, что[2]

$$\lambda_{\max}(\nabla^2 f(x)) \leqslant L,$$

т. е. все собственные значения *матрицы Гессе*

$$\nabla^2 f(x) = \left\| \frac{\partial^2 f(x)}{\partial x_i \partial x_j} \right\|_{i,j=1}^n$$

не больше L. По формуле Тейлора с остаточным членом в форме Лагранжа для любых x и y справедливо представление [93, § 58]

$$f(y) = f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\left\langle \nabla^2 f(\tilde{x})(y - x), y - x\right\rangle,$$

где $\tilde{x} = \tilde{x}(x, y)$ принадлежит отрезку, соединяющему x и y. Отсюда можно получить, что для любых x и y выполняется неравенство

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2. \tag{1.5}$$

[1] Напомним, что стационарной называют такую точку, в которой $\nabla f(x) = 0$.

[2] Строго говоря, из неравенства (1.4) выписанное неравенство следует лишь при дополнительном предположении о гладкости оптимизируемой функции. Однако основное неравенство (1.5) может быть получено и непосредственно из (1.4), см. [76, п. 1.2.2].

Из неравенства (1.5) следует, что

$$f(x^{k+1}) \leqslant f(x^k) - h\left\langle \nabla f(x^k), \nabla f(x^k)\right\rangle + \frac{Lh^2}{2}\|\nabla f(x^k)\|_2^2 = \\ = f(x^k) - h\cdot\left(1-\frac{Lh}{2}\right)\|\nabla f(x^k)\|_2^2.$$

Выбирая

$$h = \arg\max_{\alpha\geqslant 0} \alpha\cdot\left(1-\frac{L\alpha}{2}\right) = \frac{1}{L}, \tag{1.6}$$

получим

$$f(x^{k+1}) \leqslant f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|_2^2. \tag{1.7}$$

Отсюда, обозначая $x^k = x$ и учитывая, что $f(x^{k+1}) \geqslant f(x_*)$, где x_* — решение задачи (1.1), получим полезное в дальнейшем неравенство

$$\frac{1}{2L}\|\nabla f(x)\|_2^2 \leqslant f(x) - f(x_*). \tag{1.8}$$

Из неравенства (1.7) следует, что для достижения оценки

$$\min_{k=1,\ldots,N} \|\nabla f(x^k)\|_2 \leqslant \varepsilon, \tag{1.9}$$

достаточно [86, 472, 608] взять

$$N = \frac{2L\cdot(f(x^0) - f(x^{\text{extr}}))}{\varepsilon^2} \tag{1.10}$$

итераций метода (1.3) с шагом (1.6). Здесь $\nabla f(x^{\text{extr}}) = 0$ и x^{extr}, вообще говоря, зависит от точки старта x^0. Действительно, до момента выполнения неравенства (1.9) на каждой итерации согласно оценке (1.8) происходит уменьшение значения целевой функции $f(x)$ как минимум на $\varepsilon^2/(2L)$. Таким образом, не более чем после

$$N = \frac{f(x^0) - f(x^{\text{extr}})}{\varepsilon^2/(2L)}$$

итераций условие (1.9) должно выполниться первый раз.

Оценка (1.10) на классе функций, удовлетворяющих условию (1.4), с точностью до мультипликативной константы не может быть улучшена (если n не мало [191]) как для метода вида (1.3), так и для любых других методов первого порядка, т. е. использующих только градиент функции [197, 198].

♦ Здесь и далее «с точностью до мультипликативной константы» означает, что в оценке числа итераций можно попробовать улучшить числовой множитель, но не зависимость от параметров задачи. Стоит также

отметить, что сделанные оговорки «для класса функций (1.4)» и «для методов первого порядка» существенны, см., например, [195–198, 363, 472]. В частности, для достаточно гладких функций $f(x)$ (липшицев гессиан и т. д.) в классе методов первого порядка (использующих только $f(x)$ и $\nabla f(x)$) можно ожидать улучшение оценки $N \sim \varepsilon^{-2}$ до $N \sim \varepsilon^{-8/5}$ [198], а если при этом разрешается использовать в методе старшие производные оптимизируемой функции до порядка $p \geqslant 1$ включительно, то можно улучшить оценки до $N \sim \varepsilon^{-(p+1)/p}$ [172, 197]. Относительно сложности итерации таких методов при $p = 2$ см. [314, 480, 597] и цитированную там литературу. При $p \geqslant 3$ такие методы интересны в основном только в теоретическом плане. В упражнении 1.8 (п. 3) будет продемонстрировано, как можно улучшить оценку (1.10), если $f(x)$ имеет вид суммы функций.

Заметим (см. [74], а также упражнение 1.3), что для задач выпуклой оптимизации дополнительные предположения о липшицевости старших производных не меняют нижних оценок для методов первого порядка. Впрочем, если вводить более сильные предположения (самосогласованной) липшицевости (гладкости) старших производных, то и для задач выпуклой оптимизации в классе методов первого порядка возможно дополнительное ускорение [602].

Отметим также, что в общем случае для функций из класса (1.4) необходимое число итераций (на каждой итерации можно в одной и только одной точке получить значение функции $f(x)$ и её старших производных) для поиска такого x^N, что $f(x^N) - f(x_*) \leqslant \varepsilon$, где $x \in \mathbb{R}^n$, зависит от ε существенно хуже: $N \sim \varepsilon^{-n/2}$, см. [74, § 6 гл. 1]. Общая идея получения такого типа нижних оценок в наиболее простом виде изложена, например, в [76, теорема 1.1.2]. Допустимое множество разбивается на кубики со стороной $4\sqrt{2\varepsilon/L}$. Под любой метод (алгоритм) подбирается такая функция, которая во всех кубиках, кроме одного, тождественно равна нулю, а в одном кубике, том самом, который данный метод будет просматривать в последнюю очередь, функция «проваливается» (с сохранением условия (1.4)) на глубину 2ε. Поиск x^N для построенной таким образом (под рассматриваемый метод) функции потребует «просмотра» всех кубиков, число которых $\sim (1/\sqrt{\varepsilon})^n = \varepsilon^{-n/2}$.

Заметим, что далее в пособии будут приводиться примеры универсально плохих функций («худших в мире функций») для рассматриваемых классов задач сразу для всех допустимых алгоритмов, см. упражнения 1.3, 2.1 и приложение. Такую функцию можно построить и в данном случае. Действительно, рассмотрим следующую функцию Несте-

рова — Скокова (другие примеры «плохих» функций см., например, в [79]), обобщающую популярную тестовую функцию Розенброка [34, 38, 86]:

$$f(x) = \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1}(x_{i+1} - 2x_i^2 + 1)^2 =$$

$$= \frac{1}{4}(x_1 - 1)^2 + \sum_{i=1}^{n-1}\Big(x_{i+1} - \underbrace{P_2(x_i)}_{\substack{\text{многочлен}\\ \text{Чебышёва}}}\Big)^2.$$

Эта функция имеет единственный экстремум $x_* = (1, 1, \ldots, 1)$ ($f(x_*) = 0$), который и является глобальным минимумом. Если начать с точки $x^0 = (-1, 1, \ldots, 1)^T$ ($f(x^0) = 1$, $\|\nabla f(x^0)\|_2 = 1$), то метод наискорейшего спуска (см. замечание 1.4) со временем (на поздних итерациях) обеспечивает малость нормы градиента, что хорошо согласуется с описанной выше теорией, однако при этом не наблюдается сходимости по функции. Так, при $n = 16$ для градиентного спуска в момент, когда $\|\nabla f(x^N)\|_2 \approx 10^{-9}$, имеем $f(x^N) - f(x_*) \approx 0{,}98$. Похожая ситуация имеет место и для многих других популярных на практике методов, например для метода сопряжённых градиентов (см. замечание 1.6) и LBFGS (см. замечание 3 приложения). Особенность рассмотренной функции заключается в экспоненциально осциллирующих оврагах, возникающих в качестве множеств уровня функции. Связано это со следующим свойством многочленов Чебышёва:

$$P_n(P_m(x)) = P_{mn}(x)$$

и экспоненциальным ростом числа осцилляций у многочлена $P_{2^i}(x)$ с ростом i [446]. Фиксируя $x_1 = c$, можно заметить, что минимум функции $f(x)$ при $x_1 = c$ достигается в точке $x_{i+1}(c) = P_{2^i}(x_1) = P_{2^i}(c)$. Даже при близких значениях c старшие координаты $x_{i+1}(c)$ могут значительно отличаться.

Обратим внимание на то, что рассмотренные выше функции являются искусственно придуманными. Однако большие сложности, связанные с невыпуклостью (см. ниже) и многоэкстремальностью, часто возникают и в реальных приложениях, например при обучении нейронных сетей [40] или белковом фолдинге [106]. К счастью, во многих приложениях бывает достаточно найти «хороший» локальный минимум, см., например, [293, 633]. ♦

В общем случае полученный выше результат о сходимости градиентного спуска к экстремальной точке не гарантирует его сходимости

даже к локальному минимуму [76, пример 1.2.2]. Впрочем, недавно было показано [417, 418], что метод (1.3) с шагом (1.6) типично сходится именно к локальному минимуму, см. также [132]. Это по-прежнему не означает сходимость к глобальному минимуму.

Если задача (1.1) является *задачей выпуклой оптимизации*, т. е. $f(x)$ — *выпуклая функция*, что следует из неравенства $\lambda_{\min}(\nabla^2 f(x)) \geqslant 0$, то можно гарантировать сходимость метода (1.3) с шагом (1.6) к глобальному минимуму в следующем смысле [76, следствие 2.1.2]:

$$f(x^N) - f(x_*) \leqslant \frac{2LR^2}{N+4}, \tag{1.11}$$

где x_* — решение задачи (1.1), $R^2 = \|x^0 - x_*\|_2^2$. Если решение не единственно, то под x_* в формуле (1.11) можно понимать такое решение задачи (1.1), которое наиболее близко в 2-норме к точке старта x^0 [59, 76, 86].

♦ В общем случае $f(x)$ — выпуклая функция, а это означает, что её надграфик — выпуклое множество (рис. 1). Множество Q выпуклое, если вместе с любыми двумя своими точками оно содержит отрезок, их соединяющий. Это определение эквивалентно тому, что любая граничная точка множества Q отделима от этого множества, т. е. существует такая разделяющая (опорная) гиперплоскость, касающаяся множества Q в рассматриваемой точке, что множество Q лежит по одну сторону от этой гиперплоскости [66, п. 1.2, 1.3]. В таком виде далее в основном и будет использоваться понятие выпуклости — см. неравенство (1.17). Отметим, что это неравенство верно и для негладких выпуклых функций, если под $\nabla f(x)$ понимать произвольный элемент субдифференциала $\partial f(x)$ [66, п. 1.5]. ♦

Из неравенства (1.7) следует, что сходимость в смысле (1.11) гарантирует сходимость в смысле (1.9). Рассматривая функции скалярного аргумента вида $f_M(x) = x^M$, $M \gg 1$, можно заметить, что из *сходимости по функции*, т. е. в смысле (1.11), не следует в общем случае *сходимость по аргументу*. Точнее говоря, следует, но скорость сходимости по аргументу может быть сколь угодно медленной[3].

Если $f(x)$ — *μ-сильно выпуклая функция* в 2-норме, что следует из неравенства

$$\lambda_{\min}(\nabla^2 f(x)) \geqslant \mu, \quad \mu > 0,$$

[3] Ещё точнее, здесь надо ограничить класс используемых методов. При использовании методов типа деления отрезка пополам в условиях абсолютной точности вычислений возможна сходимость по аргументу и для таких (вырожденных) примеров (см. упражнение 1.4).

то для метода (1.3) с шагом (1.6) уже будет иметь место *линейная сходимость* (сходимость со скоростью геометрической прогрессии), причём по аргументу [186, теорема 3.10]:

$$\|x^N - x_*\|_2^2 \leqslant R^2 \exp\Big(-\frac{\mu}{L}N\Big), \tag{1.12}$$

где x_* — решение задачи (1.1), т. е. $\nabla f(x_*) = 0$.

Поясним, каким образом можно прийти к формуле типа (1.12). Для этого заметим, что из условия $\lambda_{\min}(\nabla^2 f(x)) \geqslant \mu$ вытекает (см. вывод неравенства (1.5) и [76, п. 2.1.3]) следующее условие для любых x и y:

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|_2^2. \tag{1.13}$$

Обычно условие (1.13) и понимают как определение μ-сильной выпуклой функции в 2-норме [76, п. 2.1.3].

Из неравенства (1.13) получается полезное в дальнейшем неравенство

$$\frac{\mu}{2}\|x - x_*\|_2^2 \leqslant f(x) - f(x_*). \tag{1.14}$$

В частности, неравенство (1.14) можно использовать для получения неравенств вида (1.12) из неравенств вида (1.16).

♦ Если неравенство (1.13) имеет место только для всех $x, y \in Q$ (или $\lambda_{\min}(\nabla^2 f(x)) \geqslant \mu$, для всех $x, y \in Q$), где Q — выпуклое множество, а $x_* = \arg\min_{x\in Q} f(x)$, то неравенство (1.14) будет иметь место для всех $x \in Q$. ♦

Из неравенства (1.13) следует, что

$$f(x_*) = \min_y f(y) \geqslant \min_y \Big\{f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|_2^2\Big\} = \\ = f(x) - \frac{1}{2\mu}\|\nabla f(x)\|_2^2,$$

т. е.

$$f(x) - f(x_*) \leqslant \frac{1}{2\mu}\|\nabla f(x)\|_2^2. \tag{1.15}$$

Отсюда с учётом неравенства (1.7) получаем

$$f(x^{k+1}) - f(x^k) \leqslant -\frac{1}{2L}\|\nabla f(x^k)\|_2^2 \leqslant -\frac{\mu}{L}\cdot(f(x^k) - f(x_*)),$$

т. е.

$$f(x^{k+1}) - f(x_*) \leqslant \Big(1 - \frac{\mu}{L}\Big)(f(x^k) - f(x_*)).$$

Следовательно,

$$f(x^N) - f(x_*) \leqslant \Big(1 - \frac{\mu}{L}\Big)^N (f(x^0) - f(x_*)) \leqslant \exp\Big(-\frac{\mu}{L}N\Big)(f(x^0) - f(x_*)). \tag{1.16}$$

♦ Если вместо настоящего градиента доступен зашумлённый градиент в концепции относительной точности[4] $\widetilde{\nabla} f(x)$ [87, п. 2, § 1; п. 3, § 2 гл. 4],

$$\|\widetilde{\nabla} f(x) - \nabla f(x)\|_2 \leqslant \alpha \|\nabla f(x)\|_2,$$

где $\alpha \in [0, 1)$, то, выбирая h в формуле (1.3) ($x^{k+1} = x^k - h\widetilde{\nabla} f(x^k)$) не по формуле (1.6), а следующим образом:

$$h = \frac{1}{L}\frac{1-\alpha}{(1+\alpha)^2},$$

вместо неравенства (1.7) получим

$$f(x^{k+1}) \leqslant f(x^k) - \frac{1}{2L}\frac{(1-\alpha)^2}{(1+\alpha)^2}\|\nabla f(x^k)\|_2^2,$$

что приведёт в итоге вместо (1.16) к неравенству

$$f(x^N) - f(x_*) \leqslant \left(1 - \frac{\mu}{L}\frac{(1-\alpha)^2}{(1+\alpha)^2}\right)^N (f(x^0) - f(x_*)).$$

Если вместо (1.15) предполагать выполнение условия (1.13), то приведённую оценку можно уточнить [227]:

$$\frac{(1-\alpha)^2}{(1+\alpha)^2} \to \mathrm{O}\left(\frac{1-\alpha}{1+\alpha}\right). \ ♦$$

Замечание 1.1 (условие градиентного доминирования). Формула (1.16) была получена в предположениях (1.4), (1.15), т. е. предположение $\lambda_{\min}(\nabla^2 f(x)) \geqslant \mu$ на самом деле использовалось лишь в виде своего следствия (1.15). Условие (1.15) называют *условием градиентного доминирования* или *условием Поляка — Лоясиевича* [376, 480]. Приведём пример, когда это условие имеет место, однако нельзя быть уверенным даже в выпуклости функции $f(x)$ [57], [75, п. 4.3], [513]. Рассмотрим систему нелинейных уравнений $g(x) = 0$, записанную в векторном виде, т. е. $g\colon \mathbb{R}^n \to \mathbb{R}^m$, $m \leqslant n$. Требуется найти какое-нибудь решение этой системы. Введём матрицу Якоби отображения g:

$$\frac{\partial g(x)}{\partial x} = \left\| \frac{\partial g_i(x)}{\partial x_j} \right\|_{i,j=1}^{m,n}.$$

Предположим, что существует такое $\mu > 0$, что для всех $x \in \mathbb{R}^n$ имеет место равномерная невырожденность матрицы Якоби:

$$\lambda_{\min}\left(\frac{\partial g(x)}{\partial x} \cdot \left[\frac{\partial g(x)}{\partial x}\right]^T\right) \geqslant \mu.$$

[4] Существуют разные концепции шума в градиенте, см., например, [86, гл. 4], [212, 221, 233]. Далее в пособии мы в основном будем работать с концепцией из [233, 235, 236], поскольку с помощью этой конструкции удобно строить универсальные методы, см. § 5.

Тогда функция $f(x) = \|g(x)\|_2^2$ удовлетворяет условию (1.15) для произвольного x_*, для которого $f(x_*) = 0$, т. е. $g(x_*) = 0$ [480].

Рассмотренная выше тестовая функция Нестерова — Скокова возникает при сведении системы нелинейных уравнений вида $x_1 = 1$, $x_2 = 2x_1^2 - 1$, …, $x_n = 2x_{n-1}^2 - 1$, к задаче оптимизации. Матрица Якоби такой системы получается нижнетреугольной, поэтому условие Поляка — Лоясиевича выполняется. Однако это не помогает численно минимизировать возникающую функцию, поскольку обусловленность данной задачи оптимизации L/μ получается экспоненциально большой по n.

Отметим также, что можно минимизировать функционал вида $f(x) = \|g(x)\|_2^2$ при условии Поляка — Лоясиевича не только градиентным методом, но и, например, методом Гаусса — Ньютона — Нестерова [75, 477]. Этот метод является методом первого порядка и имеет глобальную скорость сходимости, аналогичную скорости сходимости градиентного спуска, однако локальная скорость сходимости оказывается сверхлинейной (как у метода Ньютона, см. приложение).

Метод Гаусса — Ньютона — Нестерова неплохо себя проявил в численных экспериментах с функцией Нестерова — Скокова.

Интересный пример возникновения задач оптимизации с условием Поляка — Лоясиевича недавно был обнаружен в теории управления [279].

Многие результаты, полученные для сильно выпуклой оптимизации, могут быть естественным образом перенесены на случай выполнения условия Поляка — Лоясиевича. В качестве популярного примера упомянем здесь результаты о скорости сходимости метода стохастического градиентного спуска [116] и его варианта с редукцией дисперсии [431]. ■

Для дальнейшего построения «линейки» основных методов нам будет полезно немного по-другому посмотреть на метод градиентного спуска.

Прежде всего заметим, что если *функция* $f(x)$ *выпуклая* (см. условие (1.13) при $\mu = 0$), т. е. для любых x и y выполнено неравенство

$$f(x) + \langle \nabla f(x), y - x \rangle \leqslant f(y), \tag{1.17}$$

то $W(x) = \|x - x_*\|_2^2/2$ также будет функцией Ляпунова системы (1.2)[5].

[5] Современные приложения аппарата квадратичных функций Ляпунова для получения оценок скорости сходимости численных методов выпуклой оптимизации см., например, в работах [147, 151, 578, 579].

Действительно,

$$\frac{dW(x(t))}{dt}=\Big\langle x(t)-x_*,\frac{dx(t)}{dt}\Big\rangle=-\Big\langle\nabla f(x(t)),x(t)-x_*\Big\rangle\leqslant f(x_*)-f(x(t))\leqslant 0,$$

$$\frac{dW(x)}{dt}=0\ \Leftrightarrow\ x\in\operatorname{Arg}\min_{x\in\mathbb{R}^n}f(x).$$

Сделанное наблюдение «подсказывает» исследовать поведение последовательности $\frac{1}{2}\|x^k-x_*\|_2^2$. Согласно соотношению (1.3) имеем

$$\frac{1}{2}\|x^{k+1}-x_*\|_2^2=\frac{1}{2}\big\|(x^k-x_*)-h\nabla f(x^k)\big\|_2^2=$$

$$=\frac{1}{2}\|x^k-x_*\|_2^2-h\langle\nabla f(x^k),x^k-x_*\rangle+\frac{h^2}{2}\|\nabla f(x^k)\|_2^2. \quad (1.18)$$

Следовательно,

$$f(x^k)-f(x_*)\overset{①}{\leqslant}\langle\nabla f(x^k),x^k-x_*\rangle\overset{②}{\leqslant}$$

$$\overset{②}{\leqslant}\frac{1}{2h}\|x^k-x_*\|_2^2-\frac{1}{2h}\|x^{k+1}-x_*\|_2^2+Lh\cdot(f(x^k)-f(x^{k+1})), \quad (1.19)$$

где неравенство ① вытекает из (1.17), а неравенство ② — из равенства (1.18) и неравенства (1.7) в предположении, что $h=1/L$. Суммируя оценки (1.19) по $k=0,\ldots,N-1$ и подставляя $h=1/L$, получим

$$\sum_{k=0}^{N-1}(f(x^k)-f(x_*))\leqslant$$

$$\leqslant\frac{1}{2h}\|x^0-x_*\|_2^2-\frac{1}{2h}\|x^N-x_*\|_2^2+Lh\cdot(f(x^0)-f(x^N))\leqslant$$

$$\leqslant\frac{1}{2h}\|x^0-x_*\|_2^2+Lh\cdot(f(x^0)-f(x^N))\overset{h=1/L}{=}\frac{LR^2}{2}+f(x^0)-f(x^N).$$

Отсюда следует, что

$$\frac{1}{N}\sum_{k=1}^{N}(f(x^k)-f(x_*))\leqslant\frac{LR^2}{2N}.$$

♦ Альтернативным определением выпуклой функции служит *неравенство Иенссена*:

$$f(\alpha x+(1-\alpha)y)\leqslant\alpha f(x)+(1-\alpha)f(y)$$

для всех x и y и произвольного $\alpha\in[0,1]$, из которого по индукции вытекает следующее неравенство [182, гл. 3]:

$$f\left(\frac{1}{N}\sum_{k=1}^{N}x^k\right)\leqslant\frac{1}{N}\sum_{k=1}^{N}f(x^k).\ \blacklozenge$$

Таким образом, ввиду выпуклости функции $f(x)$ имеет место неравенство

$$f(\bar{x}^N) - f(x_*) \leqslant \frac{LR^2}{2N}, \tag{1.20}$$

где

$$\bar{x}^N = \frac{1}{N}\sum_{k=1}^{N} x^k, \tag{1.21}$$

являющееся аналогом неравенства (1.11).

Резюмируем приведённые выше результаты в немного более точной и симметричной форме [233, 249, 472, 582].

Теорема 1.1. *Пусть для численного решения задачи* (1.1)

$$f(x) \to \min_{x\in\mathbb{R}^n},$$

с функцией $f(x)$, удовлетворяющей условию (1.4), *используется градиентный спуск* (1.3), (1.6):

$$x^{k+1} = x^k - \frac{1}{L}\nabla f(x^k). \tag{1.22}$$

Тогда

$$\min_{k=1,\ldots,N} \|\nabla f(x^k)\|_2 \leqslant \sqrt{\frac{2L\cdot(f(x^0)-f(x_*))}{N}}. \tag{1.23}$$

Если дополнительно известно, что $f(x)$ — μ-сильно выпуклая функция в 2-норме, где $\mu \geqslant 0$, то[6]

$$\min_{k=1,\ldots,N} f(x^k) - f(x_*) \leqslant \frac{LR^2}{2}\min\left\{\frac{1}{N}, \exp\left(-\frac{\mu}{L}N\right)\right\}. \tag{1.24}$$

Приведённые в теореме 1.1 оценки скорости сходимости метода (1.22) точные. Немного могут быть улучшены только числовые множители [197, 198, 249, 582]. При получении оценки (1.23) (впрочем, как и оценки (1.10)) существенным образом использовалось, что рассматривается задача безусловной оптимизации (1.1) [472].

Вместо полного доказательства теоремы 1.1 ниже приводится наглядная интерпретация неравенства (1.7), лежащего в основе доказательства теоремы.

[6] Ввиду неравенства (1.7) имеем $\min_{k=1,\ldots,N} f(x^k) = f(x^N)$. Однако начиная со следующего параграфа, в котором допускается наличие неточности (2.3) и более общие способы «проектирования» (2.29) (по сравнению с обычным евклидовым), форма записи (1.24) уже будет по существу.

Замечание 1.2 («геометрия» градиентного спуска). Если понимать градиентный спуск (1.3) с шагом (1.6) следующим образом:

$$x^{k+1} = x^k - \frac{1}{L}\nabla f(x^k) = \\ = \arg\min_{x\in\mathbb{R}^n}\left\{f(x^k) + \langle \nabla f(x^k), x - x^k\rangle + \frac{L}{2}\|x - x^k\|_2^2\right\}, \quad (1.25)$$

то метод имеет естественную геометрическую интерпретацию. Параболоид вращения

$$\bar{f}_{x^k}(x) = f(x^k) + \langle \nabla f(x^k), x - x^k\rangle + \frac{L}{2}\|x - x^k\|_2^2$$

касается графика функции $f(x)$ в точке x^k и мажорирует её на всём пространстве:

$$f(x) \leqslant \bar{f}_{x^k}(x) \quad \text{для всех } x \in \mathbb{R}^n.$$

В частности,

$$f(x^{k+1}) \leqslant \bar{f}_{x^k}(x^{k+1}) = \min_{x\in\mathbb{R}^n} \bar{f}_{x^k}(x).$$

Но по построению функции $\bar{f}_{x^k}(x)$ имеем $\bar{f}_{x^k}(x^k) = f(x^k)$. Значит, переходя от точки x^k к точке минимума параболоида x^{k+1}, мы «выедаем» у функции $f(x)$ не меньше, чем у $\bar{f}_{x^k}(x)$ (рис. 2), т. е. не меньше чем

$$\bar{f}_{x^k}(x^k) - \bar{f}_{x^k}(x^{k+1}) = \frac{1}{2L}\|\nabla f(x^k)\|_2^2.$$

Таким образом можно прийти к основному соотношению (1.7).

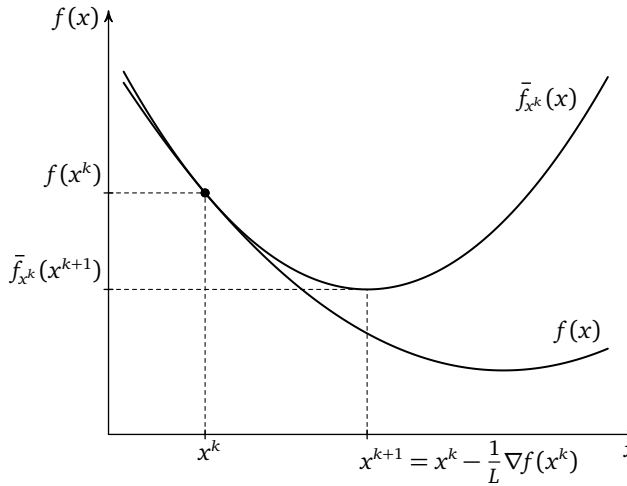

Рис. 2

Другая интерпретация имеется, например, в [79, формулы (2), (3) п. 1 § 4, гл. 1], см. также [46, п. 6.4], [56, § 5.2], [217], [495, гл. 4]. В некоторой r_k-окрестности точки x^k функция $f(x)$ заменяется линейной функцией (или более сложной моделью)

$$\widetilde{F}_{x^k}(x) = f(x^k) + \langle \nabla f(x^k), x - x^k \rangle.$$

Новое положение метода определяется исходя из решения задачи

$$x^{k+1} = \arg \min_{\|x - x^k\|_2 \leqslant r_k} \widetilde{F}_{x^k}(x).$$

С помощью *принципа множителей Лагранжа* [182, гл. 5] данная задача сводится к задаче (1.25), где $L/2$ следует понимать как множитель Лагранжа к ограничению $\|x - x^k\|_2^2 \leqslant r_k^2$. Такой подход получил название *метода доверительной области* (*trust region*). Вместе с квадратичной (ньютоновской) моделью функции его активно использовали в качестве подхода, *глобализующего сходимость* (см., например, [56, § 5.2]), т. е. обеспечивающего попадание исследуемого метода в область сверхлинейной (квадратичной) сходимости [46, п. 6.4], [217], [495, гл. 4]. Однако в последнее десятилетие данный подход в таком «глобализующем» контексте стал частично вытесняться *методом Ньютона с кубической регуляризацией* [468, 480] (см. также приложение), лучше изученным в теоретическом плане. Этот метод можно понимать как перезапись метода доверительной области с квадратичной моделью и ограничением вида $\|x - x^k\|_2^3 \leqslant r_k^3$, которое заносится в функционал с помощью принципа множителей Лагранжа, что приводит к задаче квадратичной оптимизации с дополнительным кубическим штрафным слагаемым.

Отметим также, что если рассматривается задача условной оптимизации на выпуклом множестве Q (см. § 2), то написанные выше интерпретации сохраняются. При этом в случае, когда множество Q компактно, можно выбирать, в частности, $r_k = \infty$. Получившийся в результате метод будет принадлежать к классу *методов условного градиента* [13, 86, 160, 186, 356, 403]. Получившийся метод, как и стандартный метод условного градиента (также используется название *метод Франк — Вульфа*), имеет оценки скорости сходимости на классе гладких выпуклых задач, в целом аналогичные оценкам для обычного градиентного метода [356]. Однако вместо проектирования на Q (см. § 2) на каждой итерации метода необходимо решать вспомогательную задачу минимизации линейного функционала на множестве Q. В случае, когда Q — симплекс (или шар в 1-норме), на каждой итерации получается разреженное решение вспомогательной задачи (в одной

из вершин симплекса), что позволяет существенно уменьшать стоимость итерации, см. упражнение 1.6, а также [3, 32, 186, 219].

Интересный взгляд на метод условного градиента был предложен А. С. Немировским [97], [164, п. 5.5.3], см. также [403, гл. 7]. Оказывается, такого типа методы можно также получать, взяв за основу быстрые градиентные методы в концепции модели функции (см. § 3) с неточным проектированием (см. упражнение 3.7): вместо задачи (3.3) на каждой итерации решается более простая задача — (3.3) с $1/h \equiv 0$ ($1/\alpha_{k+1} \equiv 0$ в обозначениях упражнения 3.7), и решение этой упрощённой задачи интерпретируется как приближённое решение исходной задачи (3.3). ■

Замечание 1.3 (градиентный спуск в *p*-норме). Используя приведённую выше схему рассуждений, попробуем распространить метод (1.3) на случай, когда условие (1.4) имеет более общий вид:

$$\|\nabla f(y) - \nabla f(x)\|_* \leqslant L\|y - x\|, \tag{1.26}$$

где $\|y\|_* = \max_{\|x\| \leqslant 1} \langle y, x\rangle$ — сопряжённая норма к норме $\|\ \|$. В этом случае неравенство (1.5) будет иметь аналогичный вид [127]:

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2.$$

Тогда естественно заменить метод (1.3) с шагом (1.6) следующим методом:

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ f(x^k) + \langle \nabla f(x^k), x - x^k\rangle + \frac{L}{2}\|x - x^k\|^2 \right\}. \tag{1.27}$$

Аналог неравенства (1.7) будет иметь вид [127]

$$f(x^{k+1}) \leqslant f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|_*^2.$$

Отсюда можно получить следующую оценку скорости сходимости [127]:

$$f(x^N) - f(x_*) \leqslant \frac{2L\widetilde{R}^2}{N}, \tag{1.28}$$

где $\widetilde{R} = \max_{x:\, f(x) \leqslant f(x^0)} \|x - x_*\|$. В случае $\|\ \| = \|\ \|_2$ оценку $\widetilde{R}$ можно уточнить:

$$\widetilde{R} = R = \|x^0 - x_*\|_2.$$

Если решение задачи (1.1) не единственно, то можно считать, что x_* в $\widetilde{R}^2$ выбирается таким образом, чтобы минимизировать $\widetilde{R}^2$. Оценка (1.28) внешне похожа на оценку (1.11). Однако стоит отметить, что константа L в формуле (1.28) определяется согласно неравенству (1.26), а не (1.4), и потому при $\|\ \| = \|\ \|_p$, $p \in [1, 2)$, можно ожидать, что L

в формуле (1.28) меньше, чем в формуле (1.11) [3]. Однако типично, что «выигрыш» в L с запасом нивелируется «проигрышем» в $\widetilde{R}^2$, $\widetilde{R}^2 \gg R^2$. ■

Замечание 1.4 (наискорейший спуск). Будем выбирать в методе (1.3) шаг h не из условия (1.6), а следующим образом [86, § 1, гл. 3]:

$$h^k = \arg\min_{h \geqslant 0} f(x^k - h\nabla f(x^k)). \tag{1.29}$$

Такой метод (*наискорейшего спуска*) является естественным обобщением метода градиентного спуска. Очевидно, что соотношение (1.7) сохраняется. Таким образом, можно ожидать, что *метод наискорейшего спуска* сходится не медленнее градиентного спуска. И действительно, на практике это часто можно наблюдать. Однако в худшем случае, когда[7]

$$f(x) = \frac{1}{2}\sum_{i=1}^{n} \lambda_i x_i^2, \quad 0 < \mu = \lambda_1 \leqslant \ldots \leqslant \lambda_n = L, \quad x^0 = \Big(\frac{1}{\mu}, 0, \ldots, 0, \ldots, 0, \frac{1}{L}\Big)^T,$$

наискорейший спуск сходится не быстрее градиентного спуска с постоянным (оптимально выбранным) шагом [86, теорема 3, § 4, гл. 1], вообще говоря, отличным от (1.6) [227, 582], т. е. приведённые выше оценки скорости сходимости градиентного спуска не могут быть принципиально улучшены даже при использовании шага (1.29). Рисунок 3, взятый из работы [227], демонстрирует, как сходится метод наискорейшего спуска в этом случае.

Детали можно найти, например, в работе [227]. Тем не менее движение в этом направлении (использование вспомогательной одномерной/маломерной оптимизации на каждой итерации) может приносить серьёзные дивиденды, см. замечания 1.5, 1.6.

Отметим также в этой связи следующий факт [86, § 1, гл. 3]. Если для минимизации положительно определённой квадратичной формы

$$f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle \to \min_{x \in \mathbb{R}^n} \tag{1.30}$$

использовать градиентный спуск (1.3) с $h^k = 1/\lambda_{k+1}$ где λ_{k+1} — $(k+1)$-е собственное значение матрицы A ($0 < \mu = \lambda_1 \leqslant \ldots \leqslant \lambda_n = L$), то независимо от точки старта метод будет конечен: $x^n = x_*$, где $Ax_* = b$. ■

[7] Не ограничивая общности, для задач квадратичной оптимизации можно считать, что $\lambda_1 > 0$, поскольку сильную выпуклость квадратичной задачи определяет минимальное отличное от нуля собственное значение матрицы квадратичной формы [307, 458].

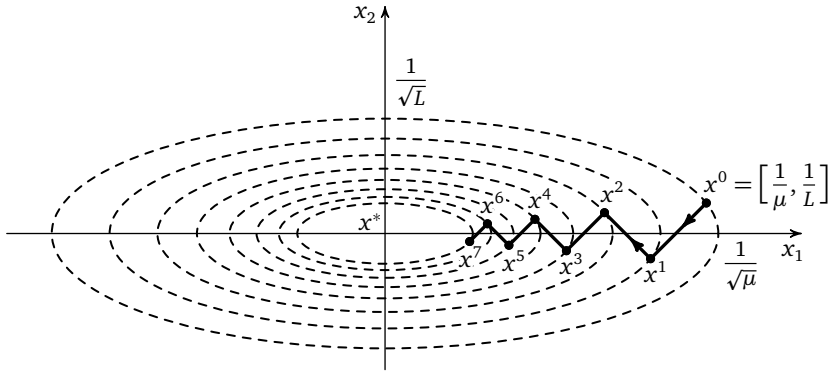

Рис. 3

Конструкции из замечаний 1.3, 1.4, к сожалению, напрямую не переносятся на обобщения, собранные в последующих параграфах. Во всяком случае, нам о такой возможности неизвестно. Однако в случае замечания 1.3 существуют «обходные пути», позволяющие за небольшую «дополнительную плату» добиться желаемого обобщения. Подробнее об этом будет написано в следующем параграфе.

Отметим также, что градиентный спуск для класса *гладких* (в смысле (1.4)) выпуклых задач оптимизации не является *оптимальным методом* (см. упражнение 1.3). Однако в следующем параграфе будет отмечено (см. упражнение 2.2), что в условиях шума градиентный спуск может оказаться оптимальным. Для класса невыпуклых гладких задач безусловной оптимизации градиентный спуск является оптимальным методом (по критерию малости 2-нормы градиента) — см. текст после формулы (1.10).

♦ Здесь и далее оптимальность метода на классе задач понимается в смысле Бахвалова — Немировского [74] — число обращений (по ходу работы метода) к оракулу за градиентом (в общем случае за старшими производными — см. замечание), т. е. число обращений к подпрограмме расчёта градиента, для достижения заданной точности (например, по функции) в зависимости от параметров, характеризующих класс рассматриваемых задач и желаемую точность, может быть уменьшено равномерно на всём рассматриваемом классе только на числовой множитель (в замечании 1.5 и у первого аргумента минимума в оценке (1.45) оптимальность понимается ещё более сильно — нельзя улучшить и числовой множитель), не зависящий от этих параметров и размерности пространства. Такой (оракульный) взгляд

на сложность задач выпуклой оптимизации оказался очень удобным и популярным [76, 186, 462]. Связано это с тем, что, с одной стороны, существует хорошо разработанная теория оракульной сложности задач выпуклой оптимизации [74], а с другой стороны, для большинства методов первого порядка и большинства задач наиболее вычислительно затратной частью итерации является именно расчёт градиента. Таким образом, число обращений к оракулу отвечает за число итераций метода, что во многом определяет и общую сложность (время работы) метода.

Книга Немировского — Юдина [74] стала в своё время (с конца 70-х годов XX века) настоящим прорывом. Эта книга во многом определила последующее развитие численных методов выпуклой оптимизации. Запас оригинальных идей, заложенных в данной книге (весьма непростой для чтения), по-прежнему вдохновляет большое число исследователей по всему миру. ♦

В заключение заметим, что естественная попытка перенести метод (1.3) на условные задачи, т. е. задачи с ограничениями $x \in Q$ простой (в смысле проектирования) структуры

$$\begin{aligned} x^{k+1} &= \arg\min_{x\in Q}\left\{\langle h\nabla f(x^k), x - x^k\rangle + \frac{1}{2}\|x - x^k\|_2^2\right\} = \\ &= \arg\min_{x\in Q}\left\{\frac{1}{2}\left\|x - (x^k - h\nabla f(x^k))\right\|_2^2 - \frac{h^2}{2}\|\nabla f(x^k)\|_2^2\right\} = \\ &= \arg\min_{x\in Q}\left\{\left\|x - (x^k - h\nabla f(x^k))\right\|_2\right\} = \pi_Q(x^k - h\nabla f(x^k)), \end{aligned} \tag{1.31}$$

приводит к аналогичному (1.18) выражению

$$\begin{aligned} \frac{1}{2}\|x^{k+1} - x_*\|_2^2 &= \frac{1}{2}\|\pi_Q(x^k - h\nabla f(x^k)) - x_*\|_2^2 = \\ &= \frac{1}{2}\left\|\pi_Q(x^k - h\nabla f(x^k) - x_*)\right\|_2^2 \leqslant \frac{1}{2}\|x^k - h\nabla f(x^k) - x_*\|_2^2 = \\ &= \frac{1}{2}\|x^k - x_*\|_2^2 - h\langle\nabla f(x^k), x^k - x_*\rangle + \frac{h^2}{2}\|\nabla f(x^k)\|_2^2. \end{aligned} \tag{1.32}$$

К сожалению, из этого неравенства уже нельзя получить неравенство (1.19), поскольку используемое при выводе (1.19) неравенство (1.7) уже может быть неверно. В следующем параграфе будет описано, как получить «правильный» аналог соотношения (1.18).

Упражнение 1.1. Докажите оценку (1.28). Объясните, почему в случае $\|\ \| = \|\ \|_2$ имеет место переход $\widetilde{R}^2 \to R^2$.

Упражнение 1.2. Докажите утверждение из последнего абзаца замечания 1.4.

Упражнение 1.3 (нижние оценки — гладкий случай / липшицев градиент). Зафиксируем N. Рассмотрим класс методов

$$x^{k+1} \in x^0 + \operatorname{Lin}\left\{\nabla f(x^0), \ldots, \nabla f(x^k)\right\}. \tag{1.33}$$

Не ограничивая общности, можно считать, что $x^0 = 0$.

1. Покажите, что в этом классе методов для *вырожденной* выпуклой функции

$$f(x) = F_{2N+1}(x) = \frac{L}{8}\left[x_1^2 + \sum_{i=1}^{2N}(x_i - x_{i+1})^2 + x_{2N+1}^2\right] - \frac{L}{4}x_1,$$

удовлетворяющей условию (1.4), при $2N+1 \leqslant n$, где $n = \dim x$, имеют место следующие нижние оценки:

$$\min_{k=1,\ldots,N} F_{2N+1}(x^k) - F_{2N+1}(x_*) \geqslant \frac{3L}{32}\frac{\|x^0 - x_*\|_2^2}{(N+1)^2},$$

$$\min_{k=1,\ldots,N} \|x^k - x_*\|_2^2 \geqslant \frac{1}{8}\|x^0 - x_*\|_2^2,$$

где

$$x_* = \arg\min_{x\in\mathbb{R}^n} F_{2N+1}(x) = \left(1 - \frac{1}{2N+2}, 1 - \frac{2}{2N+2}, \ldots, 1 - \frac{2N+1}{2N+2}, 0, \ldots, 0\right)^T.$$

2. Покажите, что в этом классе методов для μ-сильно выпуклой в 2-норме функции

$$f(x) = \frac{\mu\cdot(\chi-1)}{8}\left[x_1^2 + \sum_{i=1}^{\infty}(x_i - x_{i+1})^2 - 2x_1\right] + \frac{\mu}{2}\|x\|_2^2,$$

заданной в пространстве $\mathbb{R}^\infty$ и удовлетворяющей условию (1.4) (число обусловленности $\chi = L/\mu$), при всех $N \geqslant 1$ имеют место следующие нижние оценки:

$$f(x^N) - f(x_*) \geqslant \frac{\mu}{2}\left(\frac{\sqrt{\chi}-1}{\sqrt{\chi}+1}\right)^{2N}\|x^0 - x_*\|_2^2,$$

$$\|x^N - x_*\|_2^2 \geqslant \left(\frac{\sqrt{\chi}-1}{\sqrt{\chi}+1}\right)^{2N}\|x^0 - x_*\|_2^2.$$

♦ Заметим, что [186, теорема 3.1.5]

$$\left(\frac{\sqrt{\chi}-1}{\sqrt{\chi}+1}\right)^{2N} \simeq \exp\left(-\frac{4N}{\sqrt{\chi}}\right). \ \blacklozenge$$

Указание. См. [76, п. 2.1.2, 2.1.4], [86, § 2, гл. 3 и п. 3, § 3, гл. 12], [186, п. 3.5], [250]. Общая идея построения плохих функций под класс методов вида (1.33) следующая:

$$\textit{искать } f(x) \textit{ в виде } f(x) = \sum_{i=1}^{n-1} f_i(x_i, x_{i+1}).$$

Такая структура $f(x)$ позволяет по наперёд заданной точке x_* легко построить $f(x)$ так, чтобы минимум достигался в точке x_* [86, п. 3, § 3, гл. 12]. Однако основная цель выбора такой структуры — обеспечить при должном выборе точки старта $x^0=0$ и $f_i(x_i, x_{i+1})$ выполнение условия

$$x_i^k = 0 \quad \text{при } i > k$$

для класса методов вида (1.33). В таком случае, как бы ни выбирался метод, всегда можно, зная x_*, оценивать снизу невязку по функции. Это условие используется также при построении нижних оценок для методов первого порядка для задач выпуклой негладкой оптимизации (см. упражнение 2.1). С другой стороны, чтобы задача была сложной, необходимо, чтобы спектр гессиана $\nabla^2 f(x_*)$ был сосредоточен около максимального и минимального собственных значений (см. также замечание 1.6). При этом в выпуклом (но не сильно выпуклом) случае ещё желательно потребовать, чтобы обусловленность задачи (отношение максимального собственного значения гессиана к минимальному) была не меньше n^2 (см. методы типа центров тяжести из указания к упражнению 1.4), и необходимо потребовать, чтобы обусловленность была не меньше величины, обратной к относительной точности (по функции), с которой требуется решить задачу. Последнее условие нужно, чтобы задача была *вырожденной* [86, гл. 6]. Достаточно очевидно, что плохих функций можно подобрать много. В данном упражнении подобраны такие функции, для которых все отмеченные выше свойства достигаются на классе трёхдиагональных тёплицевых форм (а следовательно, на классе квадратичных форм с равномерно ограниченными по n коэффициентами)

$$\underbrace{\begin{pmatrix} 2 & -1 & 0 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & -1 & 2 \end{pmatrix}}_{m}.$$

Число обусловленности такой положительно определённой матрицы $\sim m^2$. Это следует из общего факта, что спектр трёхдиагональной тёплицевой матрицы, на главной диагонали которой стоит $a=2$, а на двух других $b=-1$ и $c=-1$, состоит из чисел [95, 304, 499, 585]

$$a + 2\sqrt{bc}\cos\Big(\frac{\pi k}{m+1}\Big) = 2\Big(1 - \cos\Big(\frac{\pi k}{m+1}\Big)\Big), \quad k = 1, \dots, m.$$

Отметим, что рассмотренный здесь способ порождения «плохих» функций для методов первого порядка используется практически в та-

ком же виде и для методов высокого порядка [473] (использующих старшие производные оптимизируемой функции), см. приложение. Заметим, что выписанная выше матрица является также матрицей Лапласа (Кирхгофа), см. пример 4.1.

Приведённые в условиях упражнения нижние оценки с точностью до мультипликативных констант достигаются на классе *ускоренных* (*быстрых*, *моментных*) *градиентных методов* [76, § 2.2]. За последние десять лет интерес к этому классу методов резко возрос, см., например, [20, 26, 31, 32, 40, 75, 76, 97, 120, 127, 143, 150, 151, 160, 162, 164, 186, 233, 235, 240, 241, 247, 249, 251, 252, 292, 296, 322–324, 347, 377, 385–387, 403, 433, 436–438, 457, 470, 477, 485, 492, 496, 541, 542, 544, 572, 575, 579, 582, 587, 602, 603, 608] и цитированную там литературу. Отметим также, что полученные нижние оценки сохраняют свой вид и для более общего по сравнению с (1.33) класса методов [74, гл. 7].

Приведём в простейшем случае основную идею ускорения градиентного спуска, следуя работам [127, 151]. Ограничимся только выпуклым случаем, отвечающим п. 1 упражнения 1.3. Про перенесение на сильно выпуклый случай см. конец § 5.

Начнём с неформальной идеи [127]. Рассмотрим два режима:

1) на текущей итерации $\|f'(x^k)\|_2 \geqslant M$;
2) на текущей итерации $\|f'(x^k)\|_2 < M$.

Каждый шаг (итерация) обычного градиентного спуска (1.22) в режиме 1 уменьшает невязку по функции согласно неравенству (1.7) как минимум на $M^2/(2L)$. Следовательно, верхняя оценка на число таких шагов, необходимых для достижения точности (по функции) ε, будет пропорциональна $L\varepsilon/M^2$. С другой стороны, если всё время пребывать в режиме 2 (в этом месте имеется неточность в рассуждениях!), то согласно упражнению 2.1 можно достичь точности (по функции) ε за число шагов, пропорциональное M^2/ε^2. Если выбрать параметр M так, чтобы сбалансировать обе полученные оценки: $L\varepsilon/M^2 \sim M^2/\varepsilon^2$, то общее число итераций в каждом из режимов составит $\sim \sqrt{L/\varepsilon}$, что лучше оценки $\sim L/\varepsilon$, получаемой при использовании обычного градиентного спуска.

Перейдём к более формальным выкладкам. Прежде всего заметим, что если бы в неравенстве (1.19) можно было «забыть» про необходимость выбирать шаг согласно правилу (1.6), то, выбирая $h = R/\sqrt{2L\Delta f}$, где $\Delta f = f(x^0) - f(x_*)$, $R = \|x^0 - x_*\|_2$, мы вместо (1.20) получили бы

$$f(\bar{x}^N) - f(x_*) \leqslant \frac{\sqrt{2LR^2\Delta f}}{N}, \quad \text{где} \quad \bar{x}^N = \frac{1}{N}\sum_{k=0}^{N-1} x^k.$$

Следовательно, после $N_1 = \sqrt{8LR^2/\Delta f}$ итераций мы гарантированно бы имели

$$f(\overline{x}^{N_1}) - f(x_*) \leqslant \frac{\Delta f}{2}.$$

Если *рестартовать* такую процедуру после каждого момента гарантированного уполовинивания невязки по функции, то получится метод, который достигает точности ε (по функции) после

$$N \simeq \underbrace{\sqrt{\frac{8LR^2}{\Delta f}}}_{N_1} + \underbrace{\sqrt{\frac{8LR^2}{\Delta f/2}}}_{N_2} + \ldots + \sqrt{\frac{8LR^2}{\varepsilon}} =$$

$$= \mathrm{O}\left(\sqrt{\frac{LR^2}{\varepsilon}} + \sqrt{\frac{LR^2}{2\varepsilon}} + \sqrt{\frac{LR^2}{4\varepsilon}} + \ldots\right) = \mathrm{O}\left(\sqrt{\frac{LR^2}{\varepsilon}}\right) \tag{1.34}$$

итераций. Данная оценка соответствует (с точностью до числового множителя) приведённой в условии 1 упражнения 1.3 нижней оценке. Однако всё это было получено при невыполнимом для градиентного спуска предположении, что в неравенстве (1.19) можно выбирать $h = R/\sqrt{2L\Delta f}$. Основная проблема связана с тем, что шаг h жёстко задаётся в момент использования неравенства (1.7). Однако если ввести три последовательности, связанные соотношениями ($y^0 = z^0$)

$$y^{k+1} = x^{k+1} - \frac{1}{L}\nabla f(x^{k+1}), \tag{1.35}$$

$$z^{k+1} = z^k - h\nabla f(x^{k+1}), \tag{1.36}$$

то из соотношения (1.36) получим

$$\langle h\nabla f(x^{k+1}), z^k - x_*\rangle \leqslant \frac{1}{2}\|z^k - x_*\|_2^2 - \frac{1}{2}\|z^{k+1} - x_*\|_2^2 + \frac{\|h\nabla f(x^{k+1})\|_2^2}{2},$$

а из последнего неравенства, подобно (1.19), можно получить неравенство

$$\langle \nabla f(x^{k+1}), z^k - x_*\rangle \leqslant$$
$$\leqslant \frac{1}{2h}\|z^k - x_*\|_2^2 - \frac{1}{2h}\|z^{k+1} - x_*\|_2^2 + Lh\cdot(f(x^{k+1}) - f(y^{k+1})),$$

справедливое уже для любого $h > 0$. Но, решив одну проблему, мы приобрели другую. Последнее неравенство, в отличие от (1.19), не обладает *телескопическим свойством*: при суммировании все слагаемые в правой части неравенства, кроме крайних, взаимно уничтожаются, а левая часть неравенства мажорирует невязку по функции. Чтобы добиться выполнения телескопического свойства, воспользуемся одной не использованной пока степенью свободы в определении $x^{k+1}(y^k, z^k)$.

А именно, попробуем так подобрать эту зависимость, чтобы выполнялось неравенство

$$\langle \nabla f(x^{k+1}), x^{k+1} - x_* \rangle - Lh \cdot (f(y^k) - f(y^{k+1})) \leqslant \\ \leqslant \langle \nabla f(x^{k+1}), z^k - x_* \rangle - Lh \cdot (f(x^{k+1}) - f(y^{k+1})).$$

Это удаётся сделать, используя выпуклость функции $f(x)$:

$$\langle \nabla f(x^{k+1}), y^k - x^{k+1} \rangle \leqslant f(y^k) - f(x^{k+1}),$$

если $x^{k+1} - z^k = Lh \cdot (y^k - x^{k+1})$. Получается следующая простая зависимость (*выпуклая комбинация*):

$$x^{k+1} = \tau z^k + (1-\tau) y^k, \quad \text{где} \quad \tau = \frac{1}{Lh+1}. \tag{1.37}$$

В результате для описанного здесь метода *линейного каплинга* (1.35)–(1.37) (название взято из работы [127]) можно написать следующую оценку, аналогичную (1.19):

$$\langle \nabla f(x^{k+1}), x^{k+1} - x_* \rangle \leqslant \\ \leqslant \frac{1}{2h}\|z^k - x_*\|_2^2 - \frac{1}{2h}\|z^{k+1} - x_*\|_2^2 + Lh \cdot (f(y^k) - f(y^{k+1})),$$

обладающую всеми необходимыми свойствами для последующего получения (с помощью рестартов) оптимальной оценки (1.34).

К сожалению, во-первых, описанный выше подход базируется на знании, как правило, априорно неизвестной величины R (используемой при выборе размера шага $h = R/\sqrt{2L\Delta f}$), а во-вторых, для корректности подхода под R вместо $\|x^0 - x_*\|_2$ необходимо понимать заметно бóльшую величину $\max_{x:\, f(x) \leqslant f(x^0)} \|x - x_*\|_2$, см. также замечание 1.3, в котором подобная оценка R возникает из-за использования неевклидовой нормы. Обе отмеченные проблемы (явное использование R при выборе шага и переоценка этого параметра) могут быть решены небольшой модификацией описанного подхода [127]. Кратко об этом написано в замечании 1.6 и в конце замечания 2 в приложении.

Впрочем, известно и много других вариантов ускоренных (быстрых, моментных) градиентных методов (см. начало указания), имеющих оценки глобальной скорости сходимости, аналогичные (с точностью до числового множителя) оценке (1.34), которые лишены отмеченных недостатков, см. ниже.

♦ Первым ускоренным градиентным методом с постоянными шагами для не квадратичных задач выпуклой оптимизации был (двухшаговый) *метод тяжёлого шарика*:

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta \cdot (x^k - x^{k-1}), \tag{1.38}$$

предложенный Б. Т. Поляком в 1963–1964 гг. [86, п. 1, § 2, гл. 3], [397]. Локальный анализ скорости сходимости метода тяжёлого шарика (с помощью *первого метода Ляпунова* [86, § 1, гл. 2]) при специальном выборе параметров шага $\alpha, \beta > 0$ давал правильные порядки (соответствующие нижним оценкам) локальной скорости сходимости в сильно выпуклом случае. Анализ скорости сходимости метода в среднем давал правильные порядки уже глобальной сходимости [543]. Однако с установлением глобальной сходимости (в худшем случае) были некоторые трудности. В частности, на специально подобранном примере с разрывным гессианом [347] метод может и не сходиться, а в чезаровском смысле траектории метода сходятся медленнее — аналогично (неускоренному) градиентному методу [294]. Несмотря на отмеченные сложности, метод тяжёлого шарика по-прежнему активно используется и продолжает развиваться [40, 438].

В 1982–1983 гг. Ю. Е. Нестеров в кандидатской диссертации (научным руководителем был Б. Т. Поляк) предложил первый ускоренный (быстрый) градиентный метод с фиксированными шагами (т. е. без вспомогательной маломерной оптимизации на каждой итерации, см. замечание 1.4), для которого удалось доказать глобальную сходимость с оптимальной скоростью (1.34) [77, 80]. Метод был «забыт» почти на 20 лет. Большое внимание этот метод привлёк к себе лишь после выхода в 2004 г. в издательстве Kluwer на английском языке первого издания книги [76] и работы [485]. Важную роль в привлечении внимания к методу сыграла также статья [162], имеющая большое число цитирований.

Отметим, что метод тяжёлого шарика был мотивирован овражным методом Гельфанда — Цетлина [33], [86, п. 3, § 2, гл. 6]. Ускоренный метод Нестерова также можно понимать как специальный вариант овражного метода [13, п. 3–5, § 1, гл. 5]. ♦

Для задач безусловной минимизации наиболее популярным сейчас является следующий (двухшаговый) вариант быстрого градиентного метода [76, 572] (рис. 4): $x^0 = y^0$,

$$x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k), \quad y^{k+1} = x^{k+1} + \frac{k}{k+3}(x^{k+1} - x^k),$$

который также можно понимать как *моментный метод* (следует сравнить с формулами (1.38), (1.44))

$$x^1 = x^0 - \frac{1}{L}\nabla f(x^0),$$

$$x^{k+1} = x^k - \frac{1}{L}\nabla f\Big(x^k + \frac{k-1}{k+2}(x^k - x^{k-1})\Big) + \frac{k-1}{k+2}(x^k - x^{k-1}). \qquad (1.39)$$

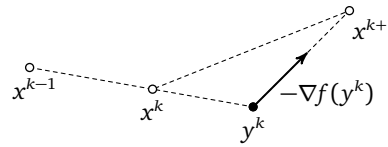

Рис. 4

Идею ускорения поясняет рис. 5, взятый из работы [533], на котором показаны линии уровня выпуклой функции и поведение траекторий градиентного спуска (слева) и быстрого градиентного метода (справа).

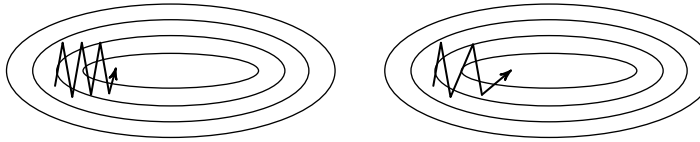

Рис. 5

Для справки приведём также вариант моментного метода для случая, когда оптимизируемая функция является μ-сильно выпуклой в 2-норме [76, (2.38) п. 2.2.1]:

$$x^1 = x^0 - \frac{1}{L}\nabla f(x^0),$$

$$x^{k+1} = x^k - \frac{1}{L}\nabla f\left(x^k + \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}(x^k - x^{k-1})\right) + \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}(x^k - x^{k-1}). \quad (1.40)$$

Оба метода (1.39), (1.40) сходятся согласно соответствующим нижним оценкам, приведённым в условиях 1, 2 упражнения 1.3, с точностью до числовых множителей при L и $\chi = L/\mu$. Более того, можно даже объединить оба метода (1.39), (1.40) в один, сходящийся (с точностью до числовых множителей при L и χ) по оценке, наилучшей из двух оценок, приведённых в условиях 1 и 2 [31], [76, теорема 2.2.2]. Однако во всех случаях (для метода (1.40) и объединённого метода) требуется априорно знать параметр μ (см. в этой связи также замечание 1.5, указания к упражнениям 2.3, 4.8 и конец § 5). На данный момент неизвестны такие варианты быстрых градиентных методов для сильно выпуклых задач, которые бы в общем случае обходились без этого знания. Отметим при этом, что незнание параметра L может быть устранено вспомогательной маломерной минимизацией (см. замечания 1.5, 1.6) или адаптивным подбором (см. § 5).

Из сопоставления соотношений (1.22) и (1.39), (1.40) легко заметить, что сложность (трудоёмкость) итераций у градиентного спуска и у быстрого градиентного метода практически одинакова, в то время как скорости сходимости отличаются очень существенно. Именно это обстоятельство и обусловило огромную популярность быстрых градиентных методов.

Замечание 1.5. Среди последних достижений в развитии ускоренных градиентных методов отметим *оптимизированный* вариант быстрого градиентного метода для задач безусловной оптимизации [247, 250, 385–387]

$$\begin{aligned} y^{k+1} &= \Big(1-\frac{1}{t_{k+1}}\Big)x^k + \frac{1}{t_{k+1}}x^0, \\ d^{k+1} &= \Big(1-\frac{1}{t_{k+1}}\Big)\nabla f(x^k) + \frac{2}{t_{k+1}}\sum_{j=0}^{k} t_j \nabla f(x^j), \\ x^{k+1} &= y^{k+1} - \frac{1}{L}d^{k+1}, \end{aligned} \tag{1.41}$$

где

$$t_{k+1} = \frac{1+\sqrt{4t_k^2+1}}{2}, \quad k = 0, \ldots, N-2,$$

$$t_N = \theta_N, \quad \theta_{k+1} = \frac{1+\sqrt{8\theta_k^2+1}}{2}, \quad k = 0, \ldots, N-1.$$

На классе гладких выпуклых задач метод сходится согласно оценке

$$f(x^N) - f(x_*) \leqslant \frac{2LR^2}{\theta_N^2}\Big(\leqslant \frac{LR^2}{N^2}\Big), \tag{1.42}$$

которая достигается, например, на функции Хьюбера

$$f(x) = \begin{cases} \dfrac{L}{2}\|x\|_2^2, & \|x\|_2 < \dfrac{R}{\theta_N^2}, \\ \dfrac{LR}{\theta_N^2}\|x\|_2 - \dfrac{LR^2}{\theta_N^4}, & \|x\|_2 \geqslant \dfrac{R}{\theta_N^2}. \end{cases} \tag{1.43}$$

Оценка (1.42) является точной оценкой скорости сходимости оптимальных методов вида (1.33) с шагами, не зависящими от оптимизируемой функции на классе гладких выпуклых задач. Другими словами, не существует такого метода вида (1.33) с фиксированными шагами, который бы на всём классе задач гладкой выпуклой оптимизации сходился по оценке, лучшей, чем (1.42). Подчеркнём, что нельзя улучшить даже числовой множитель. Из метода (1.41) можно сделать метод, не требующий знания параметра L. Для этого в процедуре (1.41)

следует заменить

$$x^{k+1} = y^{k+1} - \frac{1}{L} d^{k+1}$$

на

$$x^{k+1} = y^{k+1} - h_{k+1} d^{k+1},$$

где

$$h_{k+1} \in \operatorname{Arg}\min_{h \in \mathbb{R}} f(y^{k+1} - h d^{k+1}).$$

Такой метод также будет сходиться согласно оценке (1.42) [247]. Отметим, что *жадный метод первого порядка* (следует сравнить с (1.44))

$$x^{k+1} \in \operatorname{Arg}\min_{x \in x^0 + \operatorname{Lin}\{\nabla f(x^0), \ldots, \nabla f(x^k)\}} f(x)$$

также будет сходиться согласно оценке (1.42) и для него эта оценка также не может быть улучшена [247].

Для класса гладких сильно выпуклых задач среди методов вида (1.33) не удалось найти простое (аналитическое) описание для оптимального метода [247] (с неулучшаемым числовым множителем в оценке скорости сходимости). Наилучшие известные сейчас методы вида (1.33) с конечной памятью для данного класса задач описаны в работах [247, 544, 579].

Подобно (1.41) можно предложить (см. [247]) «универсальный» метод со вспомогательной трёхмерной оптимизацией, который также не требует на вход никаких параметров. При этом метод оптимально работает (согласно оценке (1.42)) не только на классе гладких выпуклых задач, но и на классе негладких задач. В частности, для негладких задач метод сходится согласно неулучшаемой (на классе методов вида (1.33) с фиксированными шагами) оценке [248] (следует сравнить с упражнением 2.1):

$$f(x^N) - f(x_*) \leqslant \frac{L_0 R}{\sqrt{N+1}},$$

где L_0 определяется условиями (2.4) при $\nu = 0$. Отметим, что эта оценка также возникает (и является точной) для жадного метода первого порядка, в котором субградиент $\nabla f(x^k)$ выбирается из субдифференциала $\partial f(x^k)$ таким образом, чтобы выполнялось условие $\langle \nabla f(x^k), \nabla f(x^l) \rangle = 0$, $0 \leqslant l \leqslant k-1$ [247]. К сожалению, пока непонятно, как переносить описанные выше в этом замечании методы на задачи выпуклой минимизации на множествах простой структуры. Впрочем, некоторые шаги в этом направлении уже сделаны [580, 581].

В работе [296] (см. также [470]) был предложен такой универсальный (в смысле § 5) вариант ускоренного метода, который для невыпук-

лых задач безусловной оптимизации сходится к локальному экстремуму с оптимальной скоростью (с точностью до числового множителя) по критерию малости 2-нормы градиента, а для выпуклых задач — к глобальному минимуму также с равномерно (по классам гладкости задач) оптимальной (с точностью до числового множителя) скоростью по критерию невязки по функции, см. приложение. Впрочем, в глубоком обучении (без особого теоретического обоснования — см., например, работу [521] про недостатки популярного метода Adam [230]) различные варианты быстрого градиентного метода начали использоваться заметно раньше [296], несмотря на невыпуклость возникающих там задач обучения нейронных сетей [40, 533, 575]. Отметим также эффективность быстрых градиентных методов в существенно невыпуклых задачах белкового фолдинга и докинга [106]. ■

♦ Поясним, следуя работам [247, 578, 579, 582], основную идею получения указанных в замечании 1.5 результатов. Сформулируем главную задачу: для заданного метода вида (1.33) с шагами, не зависящими от оптимизируемой функции, или жадного метода первого порядка требуется в заданном классе $F_{\mu,L}$ функций (μ-сильно выпуклых функций с L-липшицевым градиентом — все в 2-норме) при заданной точке старта x^0:

$$\alpha \cdot (f(x^0) - f(x_*)) + \beta \|\nabla f(x^0)\|_2^2 + \gamma \|x^0 - x_*\|_2^2 \leqslant C,$$

где $\alpha, \beta, \gamma \geqslant 0$, подобрать такую функцию $f \in F_{\mu,L}$, что при заданном $N \leqslant n-2$ максимально следующее выражение:

$$\tilde{\alpha} \cdot (f(x^N) - f(x_*)) + \tilde{\beta} \|\nabla f(x^N)\|_2^2 + \tilde{\gamma} \|x^N - x_*\|_2^2,$$

где $\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma} \geqslant 0$. Не ограничивая общности, можно дополнительно считать, что $x_* = 0$, $f(x_*) = 0$.

Сформулированная задача является задачей оптимизации в бесконечномерном пространстве функций из класса $F_{\mu,L}$. Пусть x^k — последовательность точек, генерируемых рассматриваемым методом, $f_k = f(x^k)$, $g^k = \nabla f(x^k)$. Будем считать, что $k \in I_N = \{0, 1, 2, \ldots, N, *\}$, где $x^* = x_* = 0$, $f_* = f(x_*) = 0$, $g^* = \nabla f(x_*) = 0$. Ключевым наблюдением является следующее утверждение.

Для набора $\{x^k, g^k, f_k\}_{k \in I_N}$ существует такая функция $f \in F_{\mu,L}$, что $f_k = f(x^k)$, $g^k = \nabla f(x^k)$, $k \in I_N$ (в этом случае будем говорить, что набор $\{x^k, g^k, f_k\}_{k \in I_N}$ интерполируется в классе функций $F_{\mu,L}$), тогда и только тогда, когда для любых $i, j \in I_N$ выполняются условия $F_{\mu,L}$-ин-

терполируемости:

$$f_i - f_j - \langle g^j, x^i - x^j\rangle \geqslant \\ \geqslant \frac{1}{2\left(1-\frac{\mu}{L}\right)}\left(\frac{1}{L}\|g^i - g^j\|_2^2 + \mu\|x^i - x^j\|_2^2 - 2\frac{\mu}{L}\langle g^i - g^j, x^i - x^j\rangle\right).$$

Доказательство данного утверждения базируется на двух вспомогательных фактах.

1. *Набор* $\{x^k, g^k, f_k\}_{k\in I_N}$ *интерполируется в классе функций* $F_{\mu,L}$ *тогда и только тогда, когда набор* $\left\{x^k, g^k - \frac{\mu}{2}\|x^k\|_2^2, f_k - \mu x^k\right\}_{k\in I_N}$ *интерполируется в классе функций* $F_{0,L-\mu}$.

Это наблюдение следует из того, что

$$f(x) \in F_{\mu,L} \iff f(x) - \frac{\mu}{2}\|x\|_2^2 \in F_{0,L-\mu}.$$

2. *Набор* $\{x^k, g^k, f_k\}_{k\in I_N}$ *интерполируется в классе функций* $F_{0,L}$ *тогда и только тогда, когда набор* $\{g^k, x^k, \langle g^k, x^k\rangle - f_k\}_{k\in I_N}$ *интерполируется в классе функций* $F_{1/L,\infty}$.

Это наблюдение следует из того, что [527, теорема 23.5]

$$f(x^k) + f^*(g^k) = \langle g^k, x^k\rangle, \quad g^k \in \partial f(x^k), \quad x^k \in \partial f^*(g^k),$$

где $f^*(g) = \sup_{x\in\mathbb{R}^n}\{\langle g, x\rangle - f(x)\}$ — сопряжённая функция к $f(x)$. При этом по теореме Фенхеля — Моро $f^{**}(x) = f(x)$ [66, п. 1.4, 2.2].

Сначала критерий $F_{\mu,L}$-интерполируемости устанавливается для класса выпуклых негладких функций $F_{0,\infty}$ (несложно понять, что это можно сделать конструктивно в классе кусочно линейных функций). В этом случае всё довольно наглядно (см. также неравенство (1.17)): для любых $i, j \in I_N$ имеем $f_i - f_j - \langle g^j, x^i - x^j\rangle \geqslant 0$. Затем к случаю $F_{0,\infty}$ последовательно, с помощью двух сделанных наблюдений, сводится и общий случай:

$$F_{\mu,L} \xrightarrow{1)} F_{0,L-\mu} \xrightarrow{2)} F_{1/(L-\mu),\infty} \xrightarrow{1)} F_{0,\infty}.$$

Рассмотрим сначала случай, когда выбран метод вида (1.33) с шагами, не зависящими от оптимизируемой функции. В этом случае можно ввести матрицу Грама $G = P^TP \succ 0$, где[8] $P = [g^0, \ldots, g^N, x^0]$. В терминах матрицы G и $f = (f_0, \ldots, f_N)^T$ задачу можно переписать

[8] Поскольку $g^k, x^0 \in \mathbb{R}^n$, ранг матрицы $G \succ 0$ не больше чем n. Если по $G = P^TP$ нужно восстановить P, то для возможности восстановления в общем случае необходимо потребовать, чтобы выполнялось условие $N + 2 \leqslant n$. В этом случае восстановить P всегда можно (причём не единственным образом), например, с помощью разложения Холецкого [95, п. 7.6].

следующим эквивалентным образом (максимизация осуществляется по G и f):

$$\langle f, b_{\tilde{\alpha}}\rangle + \langle G, \widetilde{M}_{\tilde{\beta},\tilde{\gamma}}\rangle \to \max_{\substack{f_i - f_j + \langle G, A_{ij}\rangle \leqslant 0,\ i,j\in I_N \\ \langle G, M_{\beta,\gamma}\rangle - C \leqslant 0 \\ G \succ 0}},$$

где $\langle G, M\rangle = \operatorname{tr}(GM)$ — скалярное произведение матриц, а матрицы A_{ij} определяются рассматриваемым методом вида (1.33) с шагами, не зависящими от оптимизируемой функции. Важное свойство выписанной задачи состоит в том, что это задача выпуклой оптимизации: ввиду линейности функционала эту задачу можно переписать, сохраняя структуру, и как задачу на минимум. Более того, это *задача полуопределённого программирования* (semidefinite programming), см. приложение. С помощью современных пакетов символьных вычислений для ряда конкретных методов (например, метода градиентного спуска) эту задачу вместе с двойственной к ней даже удаётся аналитически решить[9] или хотя бы построить асимптотически (по N) точное решение. Более того, в ряде случаев, базируясь на описанной технике, удаётся найти аналитическое решение и для более сложных задач, в которых метод заранее не задан и его стоит, в свою очередь, подбирать. В этом случае матрицы A_{ij} сами становятся переменными, по которым необходимо минимизировать, и задача теряет свойство выпуклости. Тем не менее даже в такой постановке иногда удаётся получить неожиданные результаты. В частности, именно таким образом был получен результат [388], описанный в замечании 5.3. Ещё более удивительной, на первый взгляд, может показаться возможность аналитически решить задачу поиска оптимального метода вида (1.3) с шагами, не зависящими от оптимизируемой функции, для класса $F_{0,L}$, где $L \leqslant \infty$, с коэффициентами в критерии $\tilde{\beta} = 0$, $\tilde{\gamma} = 0$ и в ограничении $\alpha = 0$, $\beta = 0$ [250,387]. Оказывается, в этом случае результат получается такой же, как если бы в качестве метода использовался жадный метод первого порядка [247].

[9] Отметим, что решение задачи не только даёт пример «наихудшей» функции из рассматриваемого класса для исследуемого метода, но и позволяет конструктивно показать, что полученная оценка скорости сходимости является не только нижней, но и верхней и, таким образом, точной. Для этого необходимо рассмотреть двойственную задачу. Решив двойственную задачу, можно использовать двойственные множители как веса, с которыми взвешиваются неравенства в критерии $F_{\mu,L}$-интерполируемости при конструктивном доказательстве того, что рассматриваемый метод сходится на любой функции из рассматриваемого класса $F_{\mu,L}$ не медленнее, чем предписано полученной оценкой, — равенство достигается на найденной наихудшей функции.

Рассмотрим этот метод подробнее. Ключевое наблюдение состоит в том, что такой метод генерирует последовательность $\{x^k, g^k\}_{k=0}^N$, которая определяется (однозначно, с точностью до возможного вырождения функции $f(x)$ по части аргументов) следующим образом:

$$\begin{aligned} \langle g^i, g^j \rangle &= 0, \quad 0 \leqslant j < i = 1, \dots, N \quad &| \cdot \theta_{ij}; \\ \langle g^i, x^j - x^0 \rangle &= 0, \quad 1 \leqslant j \leqslant i = 1, \dots, N \quad &| \cdot \tau_{ij}. \end{aligned}$$

Если ввести матрицу $G = P^T P \succ 0$, где[10]

$$P = [g^0, \dots, g^N, x^1 - x^0, \dots, x^N - x^0, x_* - x^0],$$

то выписанные условия, определяющие метод, вместе с условиями $F_{\mu,L}$-интерполируемости, переписанные в терминах ограничений на элементы матрицы G и значения f_k, так же как и раньше, задают выпуклую (полуопределённую) систему ограничений. Таким образом, исходная задача опять сводится к задаче выпуклой полуопределённой оптимизации. В случае $\mu = 0$, $\tilde{\beta} = 0$, $\tilde{\gamma} = 0$ и $\alpha = 0$, $\beta = 0$ эту задачу и двойственную к ней удалось аналитически решить. При этом, зная решение двойственной задачи, можно явно предъявить оптимальный метод вида (1.33) с шагами, не зависящими от оптимизируемой функции, который в худшем случае сходится не хуже[11], чем в худшем случае сходится жадный метод первого порядка. Действительно, если известно решение двойственной задачи, то известны и θ_{ij}, τ_{ij} — двойственные множители (множители Лагранжа) к ограничениям, определяющим жадный метод. Система этих ограничений может быть с помощью этих множителей эквивалентным образом переписана как

$$\left\langle g^i, \sum_{j=0}^{i-1} \theta_{ij} g^j + \sum_{j=1}^{i} \tau_{ij} \cdot (x^j - x^0) \right\rangle = 0, \quad i = 1, \dots, N.$$

Отсюда (в предположении, что $\tau_{ii} \neq 0$) получается нужное представление искомого оптимального метода вида (1.3):

$$x^k = x^0 - \sum_{j=0}^{i-1} \frac{\theta_{ij}}{\tau_{ii}} \nabla f(x^j) + \sum_{j=1}^{i-1} \frac{\tau_{ij}}{\tau_{ii}} \cdot (x^j - x^0).$$

Обеспечить выполнение выписанного условия можно и другими способами, например за счёт вспомогательной маломерной оптимизации, см. замечание 1.5.

[10] В принципе в определение матрицы P можно не включать $x_* - x^0$.

[11] В рассматриваемом здесь случае можно сказать и точнее: «сходится так же».

Критерий $F_{\mu,L}$-интерполируемости позволил получить новые интересные результаты в исследовании численных методов первого порядка с конечной памятью и шагами, не зависящими от номера итерации, в задачах (сильно) выпуклой (стохастической) оптимизации с помощью аппарата квадратичных функций Ляпунова [579] и квадратичных потенциальных функций [578]. Удалось получить необходимые и достаточные условия существования таких функций (убывающих заданным образом) и конструктивные способы их построения. В основе лежит следующее наблюдение: условия убывания функции Ляпунова, которая является потенциальной функцией на траекториях рассматриваемого метода для всех функций из рассматриваемого класса при заданных матрицах квадратичных форм, записываются как система выпуклых ограничений вида неравенств: полуопределённых ($G \succ 0$) и линейных по G, f и матрицам квадратичных форм. При этом необходимо уметь отвечать на вопрос: существуют ли такие матрицы квадратичных форм, при которых система совместна? Данная задача сводится к седловой задаче (см. § 4) правильной (выпукло/вогнутой) структуры, которая, в свою очередь, сводится к задаче выпуклого полуопределённого программирования, причём малой размерности, определяемой глубиной памяти метода. Таким образом, например, удаётся обобщить теорему Ляпунова об устойчивости линейной системы [595] на сильно выпуклые задачи и методы с конечной памятью.

Заметим также (личное сообщение А. Тейлора), что описанная выше техника может быть перенесена и на задачи, в которых вместо настоящего градиента доступен только зашумлённый градиент, например в концепции аддитивной неточности:

$$\|\widetilde{\nabla} f(x) - \nabla f(x)\|_2 \leqslant \delta$$

или относительной аддитивной неточности:

$$\|\widetilde{\nabla} f(x) - \nabla f(x)\|_2 \leqslant \alpha \|\nabla f(x)\|_2, \quad \alpha \in [0, 1).$$

Чтобы воспользоваться описанной выше техникой, стоит возвести эти неравенства в квадрат.

Насколько нам известно, это пока ещё не сделано. Впрочем, ответ на вопрос о том, как будет накапливаться неточность в оценке скорости сходимости по функции, в первом случае ожидается [147, 212, 221, 262, 307] в виде $O(\delta \|x^0 - x_*\|_2)$ (для неускоренных и ускоренных градиентных спусков, причём для ускоренных спусков требуется до-

полнительное ограничение на число итераций: $N \leqslant 2L\|x^0 - x_*\|_2/\delta)$[12], а во втором случае ответ известен только для неускоренных методов — см. текст перед замечанием 1.1. Открытым является вопрос о том, как будет накапливаться неточность для ускоренных методов с относительной аддитивной неточностью в задачах безусловной оптимизации. ♦

Замечание 1.6 (метод сопряжённых градиентов). Самой характерной задачей выпуклой оптимизации является задача минимизации положительно определённой квадратичной формы (1.30):

$$f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle \to \min_{x\in\mathbb{R}^n}.$$

Изучив данный класс задач, можно пытаться понять, как сходятся различные методы хотя бы локально (в окрестности минимума) в задачах выпуклой оптимизации. Кроме того, такие задачи возникают как вспомогательные подзадачи при использовании методов второго порядка (например, метода Ньютона, см. приложение). Известно [37, 72, 74, 76, 86, 95, 143, 495, 585], что для выпуклой задачи квадратичной оптимизации (1.30) *метод сопряжённых градиентов* (первая итерация делается согласно соотношению (1.29))

$$x^{k+1} \in \operatorname{Arg}\min_{x\in x^0+\operatorname{Lin}\{\nabla f(x^0),\ldots,\nabla f(x^k)\}} f(x) = x^k - \alpha_k\nabla f(x^k) + \beta_k\cdot(x^k - x^{k-1}), \tag{1.44}$$

где

$$(\alpha_k, \beta_k) \in \operatorname{Arg}\min_{\alpha,\beta} f\left(x^k - \alpha\nabla f(x^k) + \beta\cdot(x^k - x^{k-1})\right),$$

сходится следующим образом (причём эта оценка в общем случае не может быть улучшена по первому и второму аргументу даже в смысле

[12] Это вытекает из [262], упражнения 3.7 и следующего наблюдения: для любого $L > 0$ выполняется неравенство

$$\langle\widetilde{\nabla}f(x) - \nabla f(x), y - x\rangle \leqslant \frac{1}{2L}\|\widetilde{\nabla}f(x) - \nabla f(x)\|_2^2 + \frac{L}{2}\|y - x\|_2^2.$$

Отметим, что оценка $O(\delta\|x^0 - x_*\|_2)$ указывает на то, что аддитивная неточность в градиенте не накапливается по мере роста номера итерации. На самом деле этот факт устанавливается в предположении равномерной ограниченности последовательности, генерируемой методом. Для неускоренного градиентного спуска с неточным градиентом такое предположение можно строго установить в общем случае лишь при дополнительном условии «ранней остановки» метода (см., например, (2.18)). Отсутствие контроля за моментом остановки в общем случае может приводить к расходимости метода [511].

мультипликативного числового множителя [74, 86, 143, 463]):

$$f(x^N)-f(x_*)\leqslant \min\left\{\frac{LR^2}{2(2N+1)^2}, 2LR^2\left(\frac{\sqrt{\chi}-1}{\sqrt{\chi}+1}\right)^{2N}, \left(\frac{\lambda_{n-N+1}-\lambda_1}{\lambda_{n-N+1}+\lambda_1}\right)^2 R^2\right\}, \tag{1.45}$$

где $N\leqslant n$, $\chi=L/\mu=\lambda_n/\lambda_1$, $R^2=\|x^0-x_*\|_2^2$ и использованы обозначения из замечания 1.4. Второй аргумент в минимуме (1.45) оценивается сверху следующим образом:

$$2LR^2\exp\left(-\frac{2N}{\sqrt{\chi}}\right).$$

Полезно сопоставить эту оценку с соответствующей (сильно выпуклой) частью оценки скорости сходимости обычного градиентного спуска (1.24):

$$\left(\frac{LR^2}{2}\right)\exp\left(-\frac{N}{\chi}\right)$$

и с нижней оценкой из п. 2 упражнения 1.3. При $N=n$ метод гарантированно находит точное решение (это свойство является особенностью, отличающей методы сопряжённых градиентов от их всевозможных обобщений, см., например, [470]), что следует из последней оценки в минимуме (1.45). Сформулированный результат является фундаментальным фактом (жемчужиной) выпуклой оптимизации и вычислительной линейной алгебры одновременно [211, 245] и базируется на наличии рекуррентных формул для *многочленов Чебышёва* [74, 86, 95, 143, 304, 495, 499, 536, 541, 585, 625]. Следуя [76, п. 1.3.2], приведём рассуждения, поясняющие правое равенство в формуле (1.44). Введём обозначение (в выкладках мы воспользовались тем, что по условию $Ax_*=b$):

$$\begin{aligned}\Lambda_k=\operatorname{Lin}\left\{\nabla f(x^0),\dots,\nabla f(x^{k-1})\right\}&=\operatorname{Lin}\left\{Ax^0-b,\dots,Ax^{k-1}-b\right\}=\\&=\operatorname{Lin}\left\{A(x^0-x_*),\dots,A(x^{k-1}-x_*)\right\}.\end{aligned}$$

Заметим, что

$$\Lambda_k=\operatorname{Lin}\left\{A(x^0-x_*),\dots,A^k(x^0-x_*)\right\},$$

т. е. Λ_k есть *линейное подпространство Крылова* [74, 499]. Из определения x^{k+1} (левого равенства в формуле (1.44)) следует, что

1) для всех $p\in\Lambda_k$ выполняется равенство $\langle\nabla f(x^k),p\rangle=0$.

Введём *сопряжённые направления* $\delta^k=x^{k+1}-x^k$. Сопряжённые направления также порождают Λ_k:

2) $\Lambda_k=\operatorname{Lin}\{\delta^0,\dots,\delta^{k-1}\}$.

Название «сопряжённые направления» обусловлено следующим свойством:

3) для $k \neq i$ выполняется равенство $\langle A\delta^k, \delta^i\rangle = 0$.

Действительно, пусть для определённости $k > i$, тогда

$$\langle A\delta^k, \delta^i\rangle = \langle A(x^{k+1} - x^k), \delta^i\rangle = \left\langle \nabla f(x^{k+1}) - \nabla f(x^k), \delta^i \right\rangle \overset{1)}{=} 0.$$

Из свойства 2 и соотношения (1.44) (определения x^{k+1}) следует, что

$$x^{k+1} = x^k - h_k \nabla f(x^k) + \sum_{i=0}^{k-1} \lambda_i \delta^i,$$

т. е.

$$\delta^k = -h_k \nabla f(x^k) + \sum_{i=0}^{k-1} \lambda_i \delta^i.$$

Взяв скалярное произведение обеих частей этого равенства с вектором $A\delta^j$, по свойству 3 при $j < k-1$ получим

$$0 = \langle \delta^k, A\delta^j\rangle = -h_k \langle \nabla f(x^k), A\delta^j\rangle + \sum_{i=0}^{k-1} \lambda_i \langle \delta^i, A\delta^j\rangle =$$
$$= -h_k \underbrace{\left\langle \nabla f(x^k), \nabla f(x^{j+1}) - \nabla f(x^j)\right\rangle}_{0} + \lambda_j \langle \delta^j, A\delta^j\rangle = \lambda_j \langle \delta^j, A\delta^j\rangle,$$

т. е. $\lambda_j = 0$. Таким образом, доказано правое равенство в формуле (1.44). Оценка (1.45) получается из левого равенства (1.44) с помощью следующего наблюдения (см., например, [86, п. 2, § 2, гл. 3]): условие $x^N \in x^0 + \Lambda_N$ равносильно тому, что существует такой многочлен

$$P_N(\lambda) = 1 + a_{1N}\lambda + \ldots + a_{NN}\lambda^N,$$

где коэффициенты $a_{1N}, \ldots, a_{NN}$ могут принимать произвольные действительные значения, что $x^N - x_* = P_N(A)(x^0 - x_*)$. Поэтому для метода (1.44) должно выполняться соотношение

$$f(x^N) - f(x_*) = \frac{1}{2}\langle Ax^N, x^N\rangle - \langle b, x^N\rangle - \underbrace{\left(\frac{1}{2}\langle Ax_*, x_*\rangle - \langle b, x_*\rangle\right)}_{0} =$$
$$= \frac{1}{2}\langle A(x^N - x_*), x^N - x_*\rangle =$$
$$= \min_{a_{1N},\ldots,a_{NN}} \left\{\frac{1}{2}\left\langle AP_N(A)^2(x^0 - x_*), x^0 - x_*\right\rangle\right\} \leqslant$$
$$\leqslant \frac{1}{2} \min_{P_N(\lambda):\, P_N(0)=1} \left\{\max_{\mu=\lambda_1\leqslant\lambda\leqslant\lambda_n=L} [\lambda P_N(\lambda)^2]\right\}.$$

Таким образом, здесь появляются многочлены Чебышёва, наименее уклоняющиеся на рассматриваемом отрезке от нуля [66, п. 6.1, гл. 2]. Если относительно спектра матрицы A делать различные вероятностные предположения, то приведённый выше подход можно уточнять: вместо многочленов Чебышёва наилучшими будут уже другие многочлены [505].

Оценка (1.45) хорошо согласуется с выписанными в условии упражнения 1.3 нижними оценками. При этом оценка (1.45), полученная для класса выпуклых задач квадратичной оптимизации, лучше точной нижней оценки для общего класса задач выпуклой оптимизации (1.42).

Отметим, ввиду написанного выше, что на метод сопряжённых градиентов можно посмотреть как на наискорейший спуск (см. замечание 1.4) при выборе вместо антиградиентов сопряжённых направлений, порождаемых процедурой ортогонализации Грама — Шмидта на основе последовательности получаемых градиентов.

♦ Интересные результаты о регуляризующих свойствах метода сопряжённых градиентов для вырожденных (некорректных) задач квадратичной оптимизации были получены в цикле работ А. С. Немировского [72, 73, 463]. *Вырожденной* будем называть задачу выпуклой оптимизации, для которой отношение максимального и минимального собственного значения функционала (*обусловленность задачи*) не меньше квадрата размерности пространства, в котором происходит оптимизация: $L/\mu \gg n^2$, и не меньше величины, обратной к относительной точности, с которой требуется решить задачу, см. также указание к упражнению 1.3. Например, к такому классу задач относится задача минимизации квадратичной формы, заданной *матрицей Гильберта* [86, п. 5, § 1, гл. 11]

$$\left\| \frac{1}{i+j-1} \right\|_{i,j=1}^{n},$$

см. также приложение. Обусловленность матрицы Гильберта для $n = 6$ равна $1{,}5 \cdot 10^7$, а для $n = 10$ — $1{,}6 \cdot 10^{13}$. Многие задачи, приходящие из реальных приложений, оказываются вырожденными, см., например, [370]. Строить сходящиеся по аргументу алгоритмы для таких задач в общем случае оказывается невозможным. Решение задачи оказывается неустойчивым к неточностям в данных. Для возможности корректного восстановления решения требуются дополнительные предположения (*истокопредставимости*). Здесь мы ограничимся простейшей задачей $Ax = b$, в которой не доступны точные значения A

и b, а доступны только $\widetilde{A}$ и $\tilde{b}$, удовлетворяющие условиям

$$\|\widetilde{A}-A\|_2 \leqslant \delta_A, \quad \|\tilde{b}-b\|_2 \leqslant \delta_b,$$

где $\|C\|_2 = \sqrt{\lambda_{\max}(C^TC)}$. По поставленной задаче можно построить следующие задачи оптимизации:

1) $f_1(x) = \frac{1}{2}\|\widetilde{A}x-\tilde{b}\|_2^2 \to \min_{x\in\mathbb{R}^n}$,

2) $f_2(x) = \frac{1}{2}\langle\widetilde{A}x, x\rangle - \langle\tilde{b}, x\rangle \to \min_{x\in\mathbb{R}^n}$ (если $A^T = A \succ 0$, $\widetilde{A}^T = \widetilde{A} \succ 0$).

Введём индекс $\tau \in \{1, 2\}$, который будет отвечать рассматриваемому случаю. В работе [72] было показано, что в случае, когда выполняется *условие истокопредставимости*

$$x_* = (A^TA)^{\sigma/2}y_*, \quad \|y_*\|_2 \leqslant R_\sigma, \quad Ax_* = b,$$

метод сопряжённых градиентов с критерием останова вида

$$\|\widetilde{A}x^N-\tilde{b}\|_2 \leqslant 2(\delta_A\|x^N\|_2 + \delta_b),$$

стартующий с точки $x^0 = 0$, сходится для соответствующей задачи $\tau \in \{1, 2\}$ следующим образом:

$$\omega_N^2 = \|\widetilde{A}x^N-\tilde{b}\|_2^2 = \mathrm{O}\left(\frac{\widetilde{L}^{2(1+\sigma)}R_\sigma^2}{N^{2\tau\cdot(1+\sigma)}} + \omega_*^2\right), \quad \omega_* = \widetilde{L}^\sigma R_\sigma\delta_A + \delta_b,$$

где $\widetilde{L} = \max\{\|A\|_2, \|\widetilde{A}\|_2\}$; причём до выполнения *критерия останова*

$$\|\widetilde{A}x^N-\tilde{b}\|_2 \leqslant 2(\delta_A\|x^N\|_2 + \delta_b)$$

при $\theta + 2\sigma > 0$, $\theta \in [0, 2]$ справедлива следующая оценка:

$$\upsilon_{\theta,N}^2 = \left\|(A^TA)^{\theta/4}(x^N-x_*)\right\|_2^2 = \mathrm{O}\left(R_\sigma^{(2-\theta)/(1+\sigma)}\omega_N^{(\theta+2\sigma)/(1+\sigma)}\right),$$
$$\|x^N\|_2 = \mathrm{O}(\|x_*\|_2).$$

Обратим внимание на то, что в $\upsilon_{\theta,N}^2$ стоит настоящая (незашумлённая) матрица A. Приведённые выше результаты являются точными и не могут быть улучшены за счёт использовании других методов, причём не могут быть улучшены как в части скорости сходимости, так и в части достижимой точности $\mathrm{O}(\omega_*)$. Удивительно здесь, в частности, то, что метод сопряжённых градиентов, безусловно, можно отнести к классу ускоренных (оптимальных) методов (см. указание к упражнению 1.3), для которых известно, что в общем случае неточность в вычислении градиента линейно накапливается с ростом номера итерации [233], см. также упражнение 3.7 и начало приложения. Однако приведённый выше результат свидетельствует об отсут-

ствии накопления неточностей, что соответствует неускоренным методам [233], см. также начало приложения. Отметим при этом, что в общем случае (в отличие от рассмотренного здесь) $\widetilde{A}$, $\tilde{b}$ могут зависеть от номера итерации. Но даже в общем случае в численных экспериментах не наблюдалось накопление шума [535]. Особенно полезными приведённые выше результаты оказываются при решении некорректных обратных задач в гильбертовых пространствах [292, 370].

Важно заметить, что в описанном выше подходе не производилась регуляризация задачи, см., например, замечание 4.1 и упражнение 4.9. Оптимальный результат был достигнут за счёт регуляризирующего свойства самого метода (сопряжённых градиентов). Отметим также, что критерий останова (именно с константой 2) может быть не достижим или достижим лишь за очень длительное время, но точно достижим с некоторой другой, бóльшей (чем 2) константой. Есть некоторая проблема с правильным определением этой константы, чтобы критерий оказался достижимым за оптимальное время. Однако, как видно из приведённых выше оценок, невыполнение критерия останова не приводит к тому, что метод плохо работает. Просто, выйдя за оптимальное время на режим $\omega_N = \mathrm{O}(\omega_*)$, метод сопряжённых градиентов будет «топтаться на месте» и в итоге так и не сумеет преодолеть нужный порог. Однако, остановив его, например, по критерию превышения заданного числа итераций, гарантированно можно получить в этом случае решение с качеством

$$v_{\theta,N}^2 = \mathrm{O}\bigl(R_\sigma^{(2-\theta)/(1+\sigma)}\omega_*^{(\theta+2\sigma)/(1+\sigma)}\bigr).$$

Интересно отметить, что даже в условиях отсутствия шума $\widetilde{A} = A$, $\tilde{b} = b$ приведённые выше результаты имеют ряд неожиданных, на первый взгляд, следствий.

Предположив, что $\tau = 1$, $\sigma = 1$, $\theta = 2$, получаем оценку

$$v_{2,N}^2 = \bigl\|(A^TA)^{1/2}(x^N - x_*)\bigr\|_2^2 = \|Ax^N - b\|_2^2 = \mathrm{O}\Bigl(\frac{\widetilde{L}^4R_1^2}{N^4}\Bigr),$$

что соответствует результату, приведённому в упражнении 5.9 в условиях истокопредставимости из упражнения 4.9 с учётом отличия в обозначениях: при $\tau = 1$ имеет место равенство $\widetilde{L} \simeq \sqrt{L_{f_1}}$, где L_{f_1} — константа Липшица градиента функции $f_1(x)$.

Предположив, что $\tau = 2$, $\sigma = 0$, $\theta = 2$, получаем оценку

$$v_{2,N}^2 = \|Ax^N - b\|_2^2 = \|\nabla f_2(x^N)\|_2^2 = \mathrm{O}\Bigl(\frac{\widetilde{L}^2R_0^2}{N^4}\Bigr) = \mathrm{O}\Bigl(\frac{L_{f_2}^2\|x_*\|_2^2}{N^4}\Bigr).$$

Предположив, что $\tau = 2$, $\sigma = -1/2$, $\theta = 2$, получаем оценку

$$v_{2,N}^2 = \|Ax^N - b\|_2^2 = \|\nabla f_2(x^N)\|_2^2 = \mathrm{O}\left(\frac{\widetilde{L}R_{-1/2}^2}{N^2}\right) = \mathrm{O}\left(\frac{L_{f_2}\cdot(f_2(x_0) - f_2(x_*))}{N^2}\right).$$

Оптимальность последней оценки в контексте предпоследней кажется сомнительной, однако в замечании 5.3 будет продемонстрировано, что из последней оценки можно получить предпоследнюю как следствие.

Отметим также, что последние две оценки удаётся записать в форме, не учитывающей то, что рассматривается задача квадратичной оптимизации. Это наводит на мысль, что данные оценки должны быть справедливы для оптимальных методов и для более общего класса задач гладкой выпуклой оптимизации. Так оно и есть на самом деле. С точностью до логарифмических множителей этот результат был отмечен ещё в работе [463]. Недавно был предложен метод, который работает точно по приведённым оценкам [388], см. замечание 5.3. ♦

Метод (1.44) ничего не требует на вход (никаких параметров), а работает оптимально на классе гладких выпуклых задач и при этом также оптимально на его подклассе — классе гладких сильно выпуклых задач. Конечно, хотелось бы, чтобы и для общих задач выпуклой оптимизации метод (1.44) обладал аналогичными свойствами. Однако перенести без изменений метод (1.44) на весь класс задач выпуклой оптимизации не получилось. Тем не менее в конце 70-х годов XX века А. С. Немировскому удалось предложить две отдельные модификации метода (1.44): для класса гладких выпуклых задач и для класса гладких сильно выпуклых задач, которые доказуемо сходятся (с точностью до числовых множителей при L и χ) по оценкам, соответствующим первому аргументу минимума в оценке (1.45) и второму (в сильно выпуклом случае), см. [74, 324, 457] и цитированную там литературу. Первый метод (для выпуклых задач) также не требует на вход никаких параметров, см., например, вариант метода (1.41) с одномерной минимизацией на каждой итерации. Второй метод (для сильно выпуклых задач) использует процедуру рестартов (см. упражнение 2.3 и конец § 5) и в общем случае требует знания параметра сильной выпуклости. При этом оба метода требуют на каждой итерации решения вспомогательной малоразмерной задачи выпуклой оптимизации. В отличие от задачи (1.44), которая для квадратичных функций решается по явным формулам (см., например, [86, п. 2, § 2, гл. 3]), в общем случае на каждой итерации вспомогательную задачу можно решить только приближённо. В [74, § 3, гл. 7]

было установлено, что достаточно решать вспомогательную задачу с относительной точностью (по функции) $\delta = \mathrm{O}(\varepsilon/N(\varepsilon))$, где ε — желаемая относительная точность (по функции) решения исходной задачи, а $N(\varepsilon)$ — число итераций, которые делает (внешний) метод. Следовательно, вспомогательная задача может быть решена за

$$\mathrm{O}\Big(\ln \frac{N(\varepsilon)}{\varepsilon}\Big) = \mathrm{O}(\ln \varepsilon^{-1})$$

обращений к оракулу (подпрограмме) за значением оптимизируемой функции (см. указание к упражнению 1.4). Таким образом, оба метода получились вполне практичными. Особенно практичными эти методы оказались для задач гладкой выпуклой оптимизации с функционалом вида $f(A^T x) + g(x)$, где вычисление $A^T x$ намного дороже по времени, чем вычисление $f(y)$ и $g(x)$, см. [324, 457] и трюк из приложения с быстрым пересчётом $A^T(x + \alpha \mathrm{v})$ для разных $\alpha \in \mathbb{R}$:

$$A^T(x + \alpha \mathrm{v}) = A^T x + \alpha A^T \mathrm{v}.$$

Такие функционалы, например, возникают при решении двойственных задач к задачам минимизации выпуклых сепарабельных функционалов при аффинных ограничениях: $Ay = b$, см. § 4.

Тем не менее до сих пор так и не был найден общий метод типа (1.44) или вариант метода (1.41) с вспомогательной одномерной оптимизацией, не требующий на вход никакой информации (о гладкости / сильной выпуклости оптимизируемого функционала), который сходится оптимально (хотя бы с точностью до числовых множителей) на классе гладких выпуклых задач и при этом также оптимально на его подклассе — классе гладких сильно выпуклых задач. Мало что известно про методы типа сопряжённых градиентов (со вспомогательной маломерной оптимизацией) для задач оптимизации на множествах простой структуры [86, п. 2, § 3, гл. 7], [470]. Также мало что известно про возможные модельные обобщения (см. § 3). Лишь недавно была установлена прямодвойственность (см. § 4) специального варианта метода линейного каплинга (см. указание к упражнению 1.3), в котором вместо (1.37) осуществляется одномерная оптимизация по параметру $\tau \in [0, 1]$ [470] и(или) вместо (1.35) осуществляется одномерный поиск [322, 470]:

$$x^{k+1} = \tau_{k+1} z^k + (1 - \tau_{k+1}) y^k, \quad \tau_{k+1} \in \operatorname{Arg\,min}_{\tau \in [0,1]} f(\tau z^k + (1-\tau) y^k),$$
$$y^{k+1} = x^{k+1} - h_{k+1} \nabla f(x^{k+1}), \qquad h_{k+1} \in \operatorname{Arg\,min}_{h \geqslant 0} f(x^{k+1} - h \nabla f(x^{k+1})),$$
$$z^{k+1} = z^k - \alpha_{k+1} \nabla f(x^{k+1}),$$

где α_{k+1} определяется как решение квадратного уравнения

$$A_{k+1}f(x^{k+1}) - \frac{\alpha_{k+1}^2}{2}\|\nabla f(x^{k+1})\|_2^2 = A_{k+1}f(y^{k+1}),$$
$$A_{k+1} = A_k + \alpha_{k+1}, \quad A_0 = 0.$$

♦ Из неравенства (1.7) и того, что

$$y^{k+1} = x^{k+1} - h_{k+1}\nabla f(x^{k+1}), \quad h_{k+1} \in \operatorname{Arg}\min_{h\geqslant 0} f(x^{k+1} - h\nabla f(x^{k+1})),$$

следует, что[13]

$$\frac{\alpha_{k+1}^2}{A_{k+1}} \geqslant \frac{1}{L}, \quad A_{k+1} = A_k + \alpha_{k+1},$$

т. е.

$$\alpha_{k+1} \geqslant \frac{1}{2L} + \sqrt{\frac{1}{4L^2} + \alpha_k^2} \geqslant \frac{1}{2L} + \sqrt{\frac{1}{4L^2} + \frac{A_k}{L}}.$$

При этом

$$f(y^N) - f(x_*) \leqslant \frac{R^2}{2A_N}.$$

Далее в тексте пособия мы старались унифицировать обозначения при описании различных вариантов быстрых градиентных методов. За основу взяты обозначения из работы [127]. Обратим внимание на то, что в критерий качества в таких обозначениях следует подставлять точку y^N, а не x^N.

Важно заметить, что в таком же виде, посредством A_N, могут быть представлены оценки скорости сходимости различных ускоренных градиентных методов [32, 75, 76, 477], в том числе собранных в данном пособии. В частности, в ряде случаев на это явно указано, см. замечания 1.5, 3.3.

Если в описанном методе выбирать

$$x^{k+1} = \tau_{k+1}z^k + (1-\tau_{k+1})y^k,$$

[13] В обычных (неадаптивных) ускоренных методах это неравенство следует понимать как равенство. Таким образом, можно считать, что проблема адаптации ускоренных методов к неизвестному параметру L решается одномерным поиском (см. замечание 1.4) и формулой (1.7), из которой получается оценка неизвестного параметра

$$L = \frac{\|\nabla f(x^{k+1})\|_2^2}{2(f(x^{k+1}) - f(y^{k+1}))}.$$

Заметим, что для адаптации обычного (неускоренного) градиентного спуска достаточно только одномерного поиска, см. замечание 1.4. Заметим также, что для градиентного спуска существует вариант адаптивной оценки параметра L, не требующий вычислений значений функции [443].

где

$$\tau_{k+1} = \frac{1}{\alpha_{k+1}L} = \frac{\alpha_{k+1}}{A_{k+1}}, \quad A_{k+1} = A_k + \alpha_{k+1}, \quad A_0 = 0,$$
$$y^{k+1} = x^{k+1} - \frac{1}{L}\nabla f(x^{k+1}), \quad z^{k+1} = z^k - \alpha_{k+1}\nabla f(x^{k+1}),$$

то

$$\alpha_{k+1} = \frac{1}{2L} + \sqrt{\frac{1}{4L^2} + \alpha_k^2} = \frac{1}{2L} + \sqrt{\frac{1}{4L^2} + \frac{A_k}{L}}.$$

Полученный в результате алгоритм является обычным методом линейного каплинга (МЛК) [127], описанным в упрощённой форме в указании к упражнению 1.3. Считая, что $\alpha_k \simeq C \cdot k/L$ при $k \gg 1$, получим уравнение на C:

$$C \cdot (k+1) = \frac{1 + \sqrt{1 + 4C^2k^2}}{2},$$

следовательно,

$$1 + \frac{1}{k} = \frac{1}{2Ck} + \sqrt{1 + \frac{1}{4C^2k^2}} \simeq \frac{1}{2Ck} + 1 + \frac{1}{8C^2k^2} \simeq 1 + \frac{1}{2Ck},$$

т. е. $\alpha_k \simeq k/(2L)$, $A_k \simeq k^2/(4L)$. Следует сравнить приведённые рассуждения с немного более точным анализом, например, из работы [127], см. также конец замечания 2 в приложении. Близкий метод положен в основу проксимального ускоренного метода Монтейро — Свайтера, см. замечание 3.3. Следует также сравнить приведённые варианты МЛК с быстрым градиентным методом из упражнения 3.7 (методом подобных треугольников). ♦

Про другие методы вида (1.33) с вспомогательной маломерной оптимизацией по-прежнему ничего не известно. С учётом того, что по теоретическим оценкам использование таких методов не позволяет в общем случае улучшать даже числовые множители (см. также [247], замечание 1.4 и приложение А. С. Немировского в книге [80]), кажется, стоит оставить эти методы в стороне и спокойно двигаться дальше по направлению к методам с постоянными/заданными шагами и конечной памятью, более простым, на первый взгляд, для всестороннего теоретического анализа. В общем-то, далее в пособии реализуется именно такой план.

Однако на практике типично [38], что различные варианты метода сопряжённых градиентов (а их насчитывается уже как минимум несколько десятков [133, 282]) работают существенно быстрее ускоренных градиентных методов с постоянными шагами [377] и их адаптивных (универсальных) вариантов. На рис. 6, взятом из работы [322],

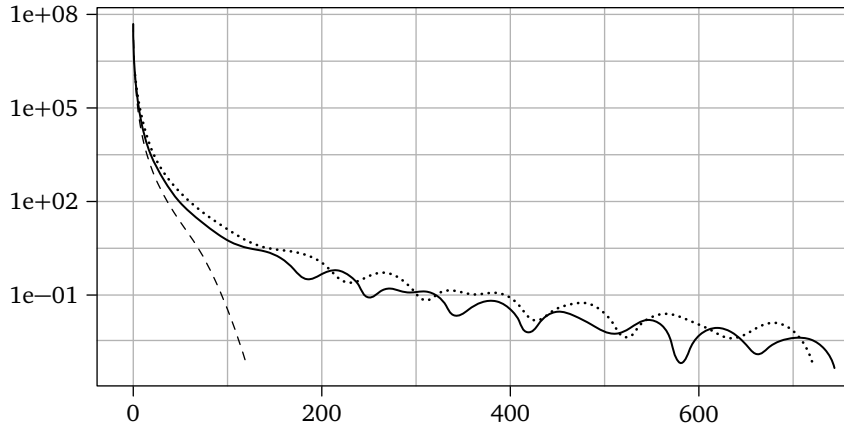

Рис. 6. На оси абсцисс откладывается число итераций, на оси ординат — невязка по функции

приведён характерный график сходимости одной из наиболее быстрых на практике версий метода сопряжённых градиентов [38, 80] (пунктирная линия) и графики сходимости быстрых градиентных методов (в данном случае использовались даже их универсальные варианты, см. § 5). Причина такого различия связана с наличием последнего аргумента минимума в оценке (1.45) для методов типа сопряжённых градиентов и со следующим наблюдением: методы типа сопряжённых градиентов сходятся так же, как ускоренные методы (с фиксированными шагами), только на специальных (как правило, малоинтересных для практики) примерах (см. замечание 1.4 и пример функции (1.43)) и на локально (в окрестности минимума) квадратичных функциях со спектром, сконцентрированным около наибольшего и наименьшего собственных значений. Например, для квадратичной формы с равномерно распределённым спектром и случайно (равновероятно) выбранной точкой старта можно ожидать (в среднем [505, 543]) сходимость специальных вариантов методов сопряжённых градиентов со скоростью (по функции) $\sim N^{-6}$ вместо ожидаемой скорости $\sim N^{-2}$ (этот результат нам сообщил Ю. Е. Нестеров). В этой связи хотелось бы обратить внимание на важность проблем, затронутых в предыдущем абзаце.

Усилить сказанное в предыдущем абзаце можно следующей цитатой из уже немного устаревшей книги [34, с. 208]: «Хотя схема сопряжённых градиентов далека от идеала, на сегодня она является единственным разумным универсальным средством решения задач

безусловной минимизации с очень большим числом переменных». Трудно сейчас всецело согласиться с такой категоричностью. Однако на практике по-прежнему именно методы типа сопряжённых градиентов и LBFGS (см. замечание 3 приложения) очень часто оказываются среди наилучших при решении выпуклых и невыпуклых задач оптимизации. Подробнее про это написано в более современной книге [38], также посвящённой практическим вопросам решения задач оптимизации больших размеров.

Отметим также в связи с рис. 6, что ускоренные градиентные методы с постоянными шагами типично сходятся не монотонно [496] (по-видимому, это можно объяснить наличием всплесков в устойчивых несимметричных линейных дискретных системах при довольно общих условиях [397, 515]), что порождает различные трудности, см., например, замечание 5.3. Впрочем, добиться монотонности только по функции (не по норме градиента и расстоянию до решения) совсем не сложно [75, п. 1.2.2], см. также упражнение 3.7. ■

♦ Наиболее популярными на практике вариантами метода сопряжённых градиентов являются следующие два метода [133], [495, гл. 5]:

$$h_k = \arg\min_{h\in\mathbb{R}} f(x^k + hp^k), \quad x^{k+1} = x^k + h_k p^k,$$
$$p^{k+1} = \nabla f(x^{k+1}) - \beta_k p^k, \quad p^0 = \nabla f(x^0),$$

$$\beta_k = -\frac{\|\nabla f(x^{k+1})\|_2^2}{\|\nabla f(x^k)\|_2^2} \quad \text{(формула Флетчера — Ривса)},$$
$$\beta_k = -\frac{\left\langle \nabla f(x^{k+1}), \nabla f(x^{k+1}) - \nabla f(x^k)\right\rangle}{\|\nabla f(x^k)\|_2^2} \quad \text{(формула Полака — Рибьера — Поляка)}.$$

Для задач квадратичной оптимизации оба метода идентичны методу (1.44). Для общих задач выпуклой оптимизации по этим методам не удалось пока получить оптимальные порядки скорости сходимости (установлен только сам факт глобальной сходимости для задач гладкой выпуклой оптимизации). Тем не менее именно эти варианты метода сопряжённых градиентов наиболее часто используются при решении практических задач [38] (в том числе не обязательно выпуклых). При этом с некоторой периодичностью (обычно период выбирают пропорционально размерности пространства, в котором происходит оптимизация) требуется перезапускать метод, обнуляя историю: вместо $p^{k+1} = \nabla f(x^{k+1}) - \beta_k p^k$ в момент рестарта полагают $p^{k+1} = \nabla f(x^{k+1})$. По-видимому, необходимость в таких рестартах

обусловлена желанием добиваться правильной сходимости в случае оптимизации сильно выпуклых функций, см. конец § 5 и [496]. ♦

Упражнение 1.4. Предложите способы решения задач выпуклой (но не обязательно сильно выпуклой) одномерной минимизации на отрезке длины Δ за время $O(\log(\Delta/\varepsilon))$, где ε — точность решения задачи по аргументу. Возможно ли такое в двумерном случае? Рассмотрите способы, базирующиеся на вычислениях значения функции и производной. Исследуйте предложенные способы на точность получаемой информации (соответственно на точность в получаемых значениях функции и в значениях её производной в разных точках). Покажите, что малейшие шумы могут привести к отсутствию сходимости в требуемую окрестность по аргументу, однако при этом можно сохранить сходимость по функции в требуемую окрестность. Покажите, что оценку можно улучшить до $O(\log\lceil\log(\Delta/\varepsilon)\rceil)$, если минимум не вырожденный (в точке минимума вторая производная положительная) и точка старта достаточно близка к точке минимума.

♦ Если ориентироваться на сходимость по функции, то для определённого класса методов (например, *метода центров тяжести*) размер области (расстояние от точки старта до решения) уже не будет входить в оценку. Будет входить лишь относительная (по функции) точность. Всё это верно лишь для сильно выпуклых задач (см., например, (1.16)) либо для выпуклых задач на выпуклых компактах с оракулом, наделённым дополнительными нетривиальными возможностями, например находить центр тяжести выпуклого компакта [186, п. 6.7]. На неограниченных множествах даже в одномерном случае в нижнюю оценку скорости сходимости по функции будет входить расстояние от точки старта до решения (в предположении, конечно, что в методе не разрешается бесплатно использовать одномерный поиск на (полу-)прямой). Причём это остаётся верным даже для выпуклых функций, имеющих ограниченную вариацию на полупрямой [74, упражнение 6 § 3, гл. 4]. К сожалению, для задач выпуклой оптимизации в $\mathbb{R}^n$, $n \gg 1$, в классе методов (1.33) даже вспомогательная маломерная оптимизация не позволяет в общем случае избавиться от вхождения расстояния от точки старта до решения в оценку скорости сходимости метода [74, 80]. Впрочем, в определённых случаях удаётся и для задач на неограниченных множествах получать оценки, в которые входит только относительная точность по функции [75, гл. 6], [477, гл. 7].

Приведённая оценка $O(\log\lceil\log(\Delta/\varepsilon)\rceil)$ как оценка скорости глобальной сходимости уже не может быть принципиально улучшена,

какие бы ни делались дополнительные предположения о гладкости функции и способностях локального оракула, вычисляющего значение функции и её старшие производные в указанной точке [74, упражнение 2 § 1, гл. 8]. В частности, для класса *чебышёвских методов* [49, п. 2.9], использующих оракулы высокого порядка, можно увеличивать основание логарифмов в рассматриваемой оценке в зависимости от свойств оптимизируемой функции и порядка оракула и тем самым улучшать оценку. Однако при этом структура оценки (повторный логарифм) останется неизменной, см. приложение, а также [144, 480]. ♦

Указание. Рассмотрите, например, метод деления отрезка пополам или метод золотого сечения [13, § 3–5, гл. 1]. Для анализа чувствительности методов и возможности ускорения при приближении к минимуму (в случае невырожденного минимума) стоит обратиться к книге [185, гл. 5].

Отметим, следуя А. С. Немировскому и Д. Б. Юдину, что для задач негладкой выпуклой, негладкой сильно выпуклой и гладкой выпуклой оптимизации на множествах простой структуры Q (см. § 2) в $\mathbb{R}^n$, когда $N \geqslant n$, нижние оценки (для более общего по сравнению с (1.33) класса методов) числа обращений к оракулу за (суб-)градиентом (что такое субградиент, будет пояснено также в § 2; про приложения можно прочитать в [451]) имеют одинаковый (с точностью до числового множителя C) вид

$$N \geqslant Cn \ln \frac{\alpha}{\varepsilon}, \tag{1.46}$$

где ε — относительная точность решения задачи по функции, а $\alpha \leqslant 1$ — отношение длин сторон вписанного и описанного параллелепипедов для Q [74]. Оценка (1.46) с $\alpha = 1$ достигается на *методе центров тяжести* Левина — Ньюмена, представляющем собой естественное обобщение на многомерные пространства метода деления отрезка пополам [74, § 3, гл. 2], [86, теорема 2, § 4, гл. 5], [186, п. 2.1].

♦ В основе метода центров тяжести лежит теорема Грюнбаума — Хаммера [66, п. 3.6], гарантирующая, что любая гиперплоскость, проходящая через центр тяжести ограниченного выпуклого множества, разделяет множество на два выпуклых подмножества, объём каждого из которых не меньше $1/e$ от объёма исходного множества. Считая, что у нас есть возможность эффективно находить центр тяжести ограниченных выпуклых множеств, на основе этой теоремы можно предложить очень простой метод: проводить через центр тяжести текущего множества, на котором происходит оптимизация, гиперплоскость,

перпендикулярную вектору (суб-)градиента, посчитанному в центре тяжести, далее отсечь от множества ту часть, в которую смотрит антиградиент. На полученном в результате множестве можно повторить описанную процедуру и т. д. После N итераций объём множества уменьшится не менее чем в $(1-1/e)^N$ раз. Выберем N так, чтобы выполнялось неравенство $(1-1/e)^N < \varepsilon^n$, т. е. $N = O(n \ln \varepsilon^{-1})$. После такого числа итераций можно быть уверенным, что найдётся хотя бы одно направление r, в проекции на которое размер множества уменьшился строго более чем в $1/\varepsilon$ раз. Отсюда следует, что значение функции в точках множества, оставшегося после N итераций, будет не больше, чем значение в точке $x_* + \varepsilon r$, не попавшей в это оставшееся множество. Поскольку оптимизируемая функция выпуклая, получаем, что

$$f(x_* + \varepsilon r) - f(x_*) \leqslant \varepsilon \cdot (f(x_* + r) - f(x_*)).$$

Отсюда следует оценка (1.46). ♦

Поскольку искать центр тяжести выпуклого множества в общем случае проблематично[14] [186, п. 6.7], у метода центров тяжести дорогая итерация [186, п. 6.7], поэтому на практике часто используют *метод эллипсоидов* [74, § 5, гл. 2], работающий по оценке[15]

$$N = O(n^2 \ln \varepsilon^{-1}),$$

но с относительно дешёвой стоимостью (трудоёмкости) итерации[16] $O(n^2)$.

♦ В основу метода эллипсоидов положено простое наблюдение: вокруг половины шара (отсечённой гиперплоскостью, проходящей через центр шара) в n-мерном пространстве за $O(n^2)$ арифметических операций можно описать (явно задав уравнением) эллипсоид (содер-

[14] Одним из лучших известных сейчас способов решения этой задачи является рандомизированный алгоритм Hit and Run, в основу которого положена идея Markov Chain Monte Carlo. Алгоритм предполагает наличие граничного оракула, который по уравнению прямой выдаёт точки пересечения этой прямой с границей множества. Время работы алгоритма оценивается как $\widetilde{O}(n^6)$.

[15] С точностью до логарифмического по n множителя в случае оптимизации на евклидово-асимметричных множествах типа шара в 1-норме пространства $\mathbb{R}^n$.

[16] В оценку этой стоимости изначально не входит расчёт (суб-)градиента. Однако обычно (суб-)градиент можно посчитать за $O(n^2)$ операций (см., например, начало § 2), поэтому можно считать, что выписанная оценка есть оценка общей сложности итерации.

жащий выбранную половину), объём которого не будет превышать

$$\left(1-\frac{1}{(n+1)^2}\right)^{n/2}$$

объёма исходного шара [164, 186, 477]. Таким образом, если изначально решение находится в явно заданном эллипсоиде, то за $O(n^2)$ арифметических операций можно так линейно преобразовать пространство, что эллипсоид преобразуется в шар. Далее через центр шара проводится гиперплоскость согласно выбранному вектору из субдифференциала целевой функции в центре шара. Затем отбрасывается та половина шара, в которую смотрит выбранный субградиент. Вокруг оставшейся половины описывается эллипсоид (см. выше), и рассуждения повторяются (по индукции). При этом если в какой-то момент центр шара окажется точкой, которая не принадлежит изначальному эллипсоиду (в момент старта), то вместо гиперплоскости, определяемой субградиентом целевой функции, в этой точке за $O(n^2)$ арифметических операций строится отделяющая гиперплоскость (эта гиперплоскость отделяет центр шара от изначального эллипсоида) и отбрасывается та половина шара, которая не пересекается с изначальным эллипсоидом. ♦

В 2015 г. был предложен метод, который с точностью до логарифмического по n множителя работает по оценке (1.46) в смысле требуемого числа итераций и по оценке[17] $\widetilde{O}(n^2)$ в смысле стоимости итерации [419]. Отметим также *метод вписанных эллипсоидов*, предложенный в 1986 г. [101, с. 253–259], и метод Вайды, предложенный в 1989 г. [186, п. 2.3], [592], проигрывающие методу 2015 г. лишь в оценке стоимости итерации. Стоимость одной итерации этих методов оценивается соответственно как $\widetilde{O}(n^{3,5})$ и $\widetilde{O}(n^{2,37})$ [419]. Отметим, что оценка $\widetilde{O}(n^{2,37})$ связана с шагом метода Ньютона[18], т. е. с решением системы линейных уравнений с симметричной матрицей, см. также

[17] К сожалению, с достаточно большим (полиномиально) логарифмическим множителем.

[18] Для вспомогательной подзадачи поиска центра *объёмного барьера*, построенного для оставшегося (к текущей итерации) от исходного множества политопа после всех отсечений (на предыдущих итерациях) гиперплоскостями полупространств. Стоит также отметить, что эта оценка представляет больше теоретический интерес, чем практический. Наилучшую практическую производительность демонстрируют методы, которые недалеко ушли в теоретическом плане от метода Штрассена с оценкой $O(n^{\log_2 7}) \gg \widetilde{O}(n^{2,37})$ [349] и разложения Холецкого $O(n^3)$ [628].

указание к упражнению 5.9 и конец приложения. Продвижение работы [419] во многом базируется на построении нового гибридного барьера и возможности эффективно осуществлять шаг метода Ньютона для минимизации такого барьера, т. е. на возможности быстро перерешивать возникающую систему линейных уравнений с симметричной матрицей: $\widetilde{\mathrm{O}}(n^{2{,}37}) \to \widetilde{\mathrm{O}}(n^2)$, см. также указание к упражнению 5.9.

За дополнительную мультипликативную плату, пропорциональную с точностью до логарифмических множителей размерности пространства n, написанное в предыдущем абзаце переносится и на безградиентные методы [88, 422] — вместо (суб-)градиента оракул выдаёт значение функции. В гладком случае это довольно очевидное утверждение, поскольку градиент можно восстановить по частным производным, каждая из которых требует расчёта значения функции в двух точках, причём одна из этих точек общая для всех компонент.

♦ Упоминая метод эллипсоидов, нельзя не отметить изящное обоснование Л. Г. Хачияном в 1978 г. с его помощью полиномиальной сложности задачи линейного программирования в битовой сложности [101, с. 453–461][19].

Предположим, что необходимо точно ответить на вопрос: совместна ли система линейных неравенств $Ax \leqslant b$ с размерами $n = \dim x$, $m = \dim b$? Будем считать, что все элементы матрицы A и вектора b являются целыми числами. Случай рациональных чисел очевидным образом сводится к целым. Если система совместна, требуется предъявить хотя бы одно точное решение x_*. Решение этой задачи (с точностью до логарифмического множителя в оценке сложности) эквивалентно точному решению задачи линейного программирования (ЛП)

$$\langle c, x\rangle \to \min_{Ax \leqslant b}$$

с целочисленными A, b и c. Напомним, что для того, чтобы найти точное решение совместной системы линейных уравнений $Ax = b$, достаточно воспользоваться алгоритмом Гаусса со сложностью $\mathrm{O}(n^3)$ арифметических операций. Аналогичный вопрос относительно системы линейных неравенств $Ax \leqslant b$ оставался открытым до конца 70-х годов прошлого века. Первый полиномиальный алгоритм был построен Л. Г. Хачияном в 1978 г. (тогда аспирантом ФУПМ МФТИ на базовой

[19] Отметим, что в общем случае поиск точного решения задачи оптимизации — NP-полная задача. Например, в упражнении 1.7 показывается, что задача поиска точного минимума выпуклого многочлена четвёртой степени на единичной сфере (не выпуклом множестве) — NP-полная задача.

кафедре ВЦ РАН) как реакция на доклад А. С. Немировского с рассказом о скорости сходимости метода эллипсоидов. В конце 70-х годов прошлого века А. С. Немировский читал в ЦЭМИ РАН курс лекций, который лёг в основу вышедшей впоследствии книги [74]. В 1982 г. Л. Г. Хачиян, А. С. Немировский и Д. Б. Юдин за цикл исследований, включающий в том числе и разработку метода эллипсоидов с доказательством полиномиальности задач ЛП в битовой сложности, были удостоены премии Фалкерсона. Опишем вкратце основной результат, полученный Л. Г. Хачияном. Положим

$$\Lambda = \sum_{i,j=1,1}^{m,n} \log_2 |a_{ij}| + \sum_{i=1}^{m} \log_2 |b_i| + \log_2(mn) + 1.$$

Первое простое, но важное наблюдение: если система $Ax \leqslant b$ совместна, то существует такой вектор x_*, что $\|x_*\|_\infty \leqslant 2^\Lambda$ и $Ax_* \leqslant b$, иначе для всех x выполнено неравенство

$$\|(Ax-b)_+\|_\infty \geqslant 2^{-(\Lambda-1)}, \quad \text{где} \quad [(z)_+]_i = \begin{cases} z_i, & z_i \geqslant 0, \\ 0, & z_i < 0. \end{cases}$$

Второе простое, но также важное наблюдение: можно переформулировать исходную задачу совместности системы $Ax \leqslant b$ как задачу негладкой выпуклой оптимизации

$$\|(Ax-b)_+\|_\infty \to \min_{\|x_*\|_\infty \leqslant 2^\Lambda},$$

которую достаточно решить с точностью $\varepsilon = 2^{-\Lambda}$. Собственно, метод эллипсоидов и был выбран в качестве алгоритма решения данной задачи ввиду геометрической скорости сходимости, т. е. зависимости числа итераций от относительной точности вида $N(\varepsilon) \sim \ln(\tilde{\varepsilon}^{-1})$. Было показано [101, с. 453–461], что, работая в $O(n\Lambda)$-битной арифметике, с затратами (оперативной) памяти $\widetilde{O}(mn + n^2)$ описанным выше образом можно найти x_* за $\widetilde{O}(n^3(n^2+m)\Lambda)$ арифметических операций.

Вопрос о полиномиальной сложности задачи ЛП в конце 70-х годов XX века стоял достаточно остро. Связано это было в первую очередь с примером Кли — Минти (1972), в котором самый популярный на тот момент метод решения задач ЛП — симплекс-метод, прежде чем найти решение, последовательно проходит через все 2^m вершины гиперкуба, отвечающего m специально сконструированным аффинным ограничениям [464]. Только в первой половине 80-х годов удалось объяснить хорошую работу симплекс-метода на практике. Оказалось, что математическое ожидание времени работы симплекс-метода

при некоторых вполне естественных способах задания распределения вероятностей на параметрах задачи ЛП можно оценить следующим образом: $\widetilde{\mathrm{O}}(m^3)$ [15, 176, 558].

В начале этого столетия Д. Спилманом и А. Тенгом был доказан более тонкий результат [563]: если матрица аффинных ограничений в задаче ЛП имеет вид $A + \|A\|G$, где матрица $G = \|g_{ij}\|_{i,j=1,1}^{m,n}$ состоит из независимых одинаково распределённых нормальных случайных величин $g_{ij} \in N(0, \sigma^2)$ и $T_\sigma(A)$ — время работы специальной версии симплекс-метода, необходимое для нахождения точного решения, то [222]

$$E_G[T_\sigma(A)] = \mathrm{O}\bigl(n^2\sqrt{\ln m}\cdot\sigma^{-2} + n^3\ln^{3/2} m\bigr).$$

Можно ли в этой формуле (или её аналогах) для математического ожидания времени работы заменить σ^{-2} на $\log^r(\sigma^{-1})$, насколько нам известно, до сих пор открытый вопрос.

Рекордные результаты по сложности решения описанной задачи ЛП таковы: число итераций специального варианта метода внутренней точки составляет $\widetilde{\mathrm{O}}(\sqrt{\operatorname{rank} A})$, см. [420], указание к упражнению 5.9 и замечание 4 приложения. На каждой итерации необходимо решать $\widetilde{\mathrm{O}}(1)$ систем линейных уравнений, что может быть сделано за время $\widetilde{\mathrm{O}}(nnz(A))$, см. [419], указание к упражнению 5.9 и замечание 4 приложения. Однако на данный момент эти результаты представляют в основном только теоретический интерес (методы не практичные).

Отметим, что вопрос о полиномиальной сложности задач ЛП в идеальной арифметике (в которой допустимы любые числа и все арифметические операции, в частности умножение $\pi \cdot e$, выполняются за $\mathrm{O}(1)$ операций) по-прежнему остаётся открытым и был выделен С. Смейлом в качестве одной из главных математических проблем этого столетия [174].

Отметим, что на практике для решения задач небольшого размера ($n \sim 10^2$) часто используют *bundle method*[20] (см. [164, п. 5.4], [424]) и вариации метода эллипсоидов (*методы с процедурой растяжения пространства*), восходящие к работам Н. З. Шора [86, § 4, гл. 5], [90, 104]. Стоит также отметить большую роль, которую сыграли работы Н. З. Шора в появлении *субградиентного метода* (см. § 2). ♦

В случае, когда $N \leqslant n$ (обычно это соответствует задачам оптимизации в пространстве большой размерности), и при определённых

[20] Названный в работе [76, п. 3.3.3] методом уровней. Однако в данном пособии под методом уровней понимается немного другой метод, см. пример 3.2.

условиях в гладком сильно выпуклом случае оценка (1.46) перестаёт быть оптимальной (точной нижней границей сложности) [74]. Оптимальные оценки в этом случае будут достигаться на методах типа градиентного спуска, см., например, [76, 186], указание к упражнению 1.3 и замечания 1.5, 1.6.

Упражнение 1.5 (Ю. Е. Нестеров — Д. А. Пасечнюк — Ф. С. Стонякин, 2018 [504]). Рассмотрим следующий метод решения задачи минимизации выпуклой липшицевой функции (с константой Липшица L_0) на квадрате в $\mathbb{R}^2$ со стороной R. Через центр квадрата проводится горизонтальная прямая. На отрезке, высекаемом из квадрата этой прямой, с точностью $\sim \varepsilon/\log(L_0R/\varepsilon)$ (по функции) решается задача одномерной оптимизации. В найденной точке вычисляется вектор (суб-)градиента функции и определяется, в какой из двух прямоугольников он «смотрит»; этот прямоугольник «отбрасывается». Через центр оставшегося прямоугольника проводится вертикальная прямая, и на отрезке, высекаемом этой прямой в прямоугольнике, также с точностью $\sim \varepsilon/\log(L_0R/\varepsilon)$ (по функции) решается задача одномерной оптимизации. В найденной точке вычисляется вектор (суб-)градиента функции и определяется, в какой из двух квадратов он «смотрит»; этот квадрат «отбрасывается». В результате такой процедуры линейный размер исходного квадрата уменьшается вдвое.

1. Покажите, что если оптимизируемая функция гладкая (дифференцируемая), то после $\sim\log(L_0R/\varepsilon)$ повторений такой процедуры можно найти с точностью ε (по функции) решение исходной задачи.
2. Покажите, что для негладкой выпуклой функции такой метод может не сходиться к решению задачи даже по функции.

   Указание. Рассмотрите функцию $|x-y|+0{,}9x$ на $[0,1]^2$.

Упражнение 1.6 (метод условного градиента для задач квадратичной оптимизации на симплексе [32, п. 4.2.2, 4.3.3]). Рассмотрим задачу квадратичной выпуклой оптимизации

$$f(x)=\frac{1}{2}\langle Ax,x\rangle \to \min_{x\in S_n(1)},$$

где все элементы матрицы $A\succ 0$ по модулю не больше M и число ненулевых элементов в каждом столбце (строке) матрицы A не больше $s\ll n$. Для решения этой задачи будем использовать метод условного градиента (см. замечание 1.2). Выберем одну из вершин симплекса и возьмём точку старта x^0 в этой вершине. Далее действуем по индук-

ции, шаг которой имеет следующий вид. Решаем задачу

$$\langle \nabla f(x^k), y\rangle = \langle Ax^k, y\rangle \to \min_{y\in S_n(1)}.$$

Обозначим решение этой задачи через

$$y^k = (0,\dots,0,1,0,\dots,0),$$

где 1 стоит на позиции

$$i_k \in \operatorname{Arg}\min_{i=1,\dots,n} \frac{\partial f(x^k)}{\partial x_i}.$$

Несложно показать, что решение такого вида всегда есть. Далее положим

$$x^{k+1} = (1-\gamma_k)x^k + \gamma_k y^k, \quad \gamma_k = \frac{2}{k+2}.$$

Заметим, что в такой метод не входят никакие параметры!

Имеет место следующая оценка скорости сходимости описанного метода:

$$f(x^N) - f(x_*) \leqslant \frac{2L^p R_p^2}{N},$$

где

$$R_p^2 = \max_{x,y\in S_n(1)} \|y-x\|_p^2, \quad L^p = \max_{\|h\|_p\leqslant 1} \langle h, Ah\rangle, \quad 1 \leqslant p \leqslant \infty,$$

причём p тут можно выбирать произвольно. С учётом того, что оптимизация происходит на симплексе, выберем $p=1$. Несложно показать, что этот выбор оптимален. В результате получим, что $R_1^2 = 2$,

$$L^1 = \max_{i,j=1,\dots,n} |A_{ij}| \leqslant M.$$

Покажите, что после предварительных приготовлений (*препроцессинга*), имеющих сложность $\mathrm{O}(n)$, каждую итерацию можно осуществлять с трудоёмкостью $\mathrm{O}(s\log_2 n)$. Таким образом, общая трудоёмкость метода будет составлять

$$\mathrm{O}\left(n + \frac{M}{\varepsilon} s\log_2 n\right),$$

что может быть значительно лучше оценки

$$\mathrm{O}\left(sn\sqrt{\frac{M\ln n}{\varepsilon}}\right) = \underbrace{\mathrm{O}(sn)}_{\substack{\text{сложность}\\ \text{итерации}}} \cdot \underbrace{\mathrm{O}\left(\sqrt{\frac{M\ln n}{\varepsilon}}\right)}_{\text{число итераций}},$$

которая получается при использовании быстрого градиентного метода (оптимального по числу обращений к оракулу за градиентом) с наилучшей для данной задачи прокс-структурой (из известных) — энтропийной (см. конец § 2 и упражнение 3.7).

♦ Методы условного градиента можно при должной модификации применять к задачам сильно выпуклой оптимизации (см. [194, 381, 401, 403] и цитированную там литературу) и к выпукло-вогнутым седловым задачам (см. [298, 403, 411] и цитированную там литературу). Недавно появились тензорные методы условного градиента [242]. Используя нижние оценки для методов типа условного градиента (методов с линейным минимизационным оракулом) в не сильно выпуклом случае [403, п. 7.1.3] и конструкцию каталист (см. замечание 3.3 и приложение), можно получить нижние оценки в сильно выпуклом случае [108]. Эти оценки показывают, что, в отличие от выпуклого случая, в сильно выпуклом случае методы условного градиента сходятся существенно медленнее. Попытки исправить ситуацию требуют дополнительных предположений относительно возможностей линейного минимизационного оракула [403, п. 7.1.1.4]. С другими возможностями и обобщениями методов условного градиента можно познакомиться по работе [403, гл. 7]. ♦

Упражнение 1.7 (Ю. Е. Нестеров [78]). Пусть задан набор неотрицательных целых чисел $\{a_i\}_{i=1}^n$, полиномиально по n неограниченных. Покажите, что следующие задачи NP-полны [290].

1. Минимизация (точный поиск минимума) многочлена четвёртой степени

$$f(x) = \sum_{i=1}^n x_i^4 - \frac{1}{n}\left(\sum_{i=1}^n x_i^2\right)^2 + \left(\sum_{i=1}^n a_i x_i\right)^4 + (1-x_1)^4.$$

Эта задача эквивалентна задаче минимизации выпуклого многочлена четвёртой степени

$$P_4(x) = \sum_{i=1}^n x_i^4 + \left(\sum_{i=1}^n a_i x_i\right)^4 + (1-x_1)^4$$

на сфере.

2. Задача оптимального управления

$$P_4(x(1)) \to \min_{u(\,)}, \quad \frac{dx}{dt} = \frac{\langle x, u\rangle}{\|x\|_2^2}x - u, \quad 0 \leqslant t \leqslant 1,$$

где вектор $x(0) = x_0 \in \mathbb{R}^n$ задан, причём $\|x_0\|_2 = 1$. Требуется найти точное решение.

3. Поиск направления убывания невыпуклой негладкой функции

$$f(x) = \left(1 - \frac{1}{\gamma}\right)\max_{i=1,\ldots,n}|x_i| - \min_{i=1,\ldots,n}|x_i| + |\langle a, x\rangle|,$$

где $\gamma = \sum_{i=1}^n a_i > 1$, в точке $x = 0$. Нужно найти хотя бы один такой вектор $x \in \mathbb{R}^n$, что $f(x) < f(0) = 0$.

Указания. 1. Заметим, что

$$\sum_{i=1}^n x_i^4 - \frac{1}{n}\left(\sum_{i=1}^n x_i^2\right)^2 = \langle B[x]^2, [x]^2\rangle \geqslant 0,$$

где

$$[x]^2]_i = x_i^2, \quad B = I - \frac{1}{n} 1_n 1_n^T \succ 0,$$

причём $B[x]^2 = 0$ тогда и только тогда, когда $[x]^2 = \mathrm{const} \cdot 1_n$. Значит, поиск такого $x \in \mathbb{R}^n$, что

$$f(x) = \min_{x \in \mathbb{R}^n} f(x) = 0,$$

равносилен поиску такого x, что

$$(1 - x_1)^4 = 0, \quad [x]^2 = \mathrm{const} \cdot 1_n, \quad \left(\sum_{i=1}^n a_i x_i\right)^4 = 0,$$

что эквивалентно решению *задачи о рюкзаке* (*ранце*)

$$a_1 + \sum_{i=2}^n a_i x_i = 0, \quad x_i = \pm 1.$$

Таким образом, если бы можно было эффективно точно решить исходную задачу оптимизации, то можно было бы эффективно решить и задачу о рюкзаке. С помощью *быстрого преобразования Фурье* можно решить задачу о рюкзаке за время

$$\mathrm{O}\left(\ln n \cdot \sum_{i=1}^n |a_i|\right).$$

Однако по условию задачи мы не можем считать эту оценку полиномиальной, т. е. мы не предполагаем, что выполняются условия, позволяющие убрать NP-полную задачу о рюкзаке из класса NP-полных задач[21].

2. Заметим, что

$$\frac{d\|x\|_2^2}{dt} = \left\langle 2x, \frac{\langle x, u\rangle}{\|x\|_2^2} x - u \right\rangle = 2\left\langle x, \left(\frac{xx^T}{\|x\|_2^2} - I\right) u \right\rangle \equiv 0,$$

т. е. $x(1)$ принадлежит единичной сфере, как и $x(0)$. Более того, произвола в выборе $\{u(t)\}_{t \in [0,1]}$ достаточно, чтобы получить в качестве $x(1)$

[21] Другой пример, подобный рассмотренному, см. в докладе С. Сра [632].

произвольную точку на единичной сфере. Таким образом, исходная постановка задачи сводится к задаче, рассмотренной в п. 1.

3. Сначала заметим, что если набор $\sigma_i = \pm 1$ удовлетворяет условию $\langle a, \sigma\rangle = 0$, то $f(\sigma) = -1\gamma < 0$.

Пусть $f(x) < 0$. Ввиду линейной однородности $f(x)$ можно считать, не ограничивая общности, что $\max_{i=1,\dots,n} |x_i| = 1$. Обозначим $\delta = \langle a, \sigma\rangle$. Тогда $|x_i| > 1 - 1/\gamma + \delta$, $i = 1, \dots, n$. Вводя $\sigma_i = \operatorname{sign} x_i$, получим $\sigma_i x_i > 1 - 1/\gamma + \delta$, следовательно, $|\sigma_i - x_i| = 1 - \sigma_i x_i < 1/\gamma - \delta$. Поэтому

$$|\langle a, \sigma\rangle| \leqslant |\langle a, x\rangle| + |\langle a, \sigma - x\rangle| \leqslant \delta + \gamma \max_{i=1,\dots,n} |\sigma_i - x_i| < (1-\gamma)\delta + 1 < 1.$$

Поскольку по предположению вектор a имеет целочисленные компоненты, последнее неравенство означает просто, что $\langle a, \sigma\rangle = 0$.

Упражнение 1.8. 1. Рассмотрим задачу невыпуклой оптимизации

$$f(x) \to \min_{x \in \mathbb{R}^n}$$

в условиях L-липшицева градиента (1.4) и метод градиентного спуска с переменным шагом $h_k \leqslant 1/(2L)$ и неточным градиентом $\widetilde{\nabla} f(x)$:

$$x^{k+1} = x^k - h_k \widetilde{\nabla} f(x^k), \quad \text{если} \quad \|\widetilde{\nabla} f(x^k)\|_2^2 > \frac{2\varepsilon^2}{5};$$

$$\text{STOP} \quad \text{иначе},$$

где

$$\|\widetilde{\nabla} f(x) - \nabla f(x)\|_2^2 \leqslant \frac{\varepsilon^2}{10}.$$

Покажите, что такой метод гарантированно остановится не более чем за

$$N = \frac{4(f(x^0) - f(x_*))}{\varepsilon^2 \varsigma}$$

итераций, см. (1.10), где

$$\varsigma = \min_{k=1,\dots,\text{STOP}} \frac{\|\widetilde{\nabla} f(x^k)\|_2^2 h_k}{\varepsilon^2}.$$

Таким образом, можно гарантировать, что в момент остановки выполнено неравенство

$$\|\nabla f(x^k)\|_2^2 = \left\|\nabla f(x^k) - \widetilde{\nabla} f(x^k) + \widetilde{\nabla} f(x^k)\right\|_2^2 \leqslant$$
$$\leqslant 2\|\widetilde{\nabla} f(x) - \nabla f(x)\|_2^2 + 2\|\widetilde{\nabla} f(x^k)\|_2^2 \leqslant \varepsilon^2.$$

♦ В следующих двух пунктах для наглядности вместо неравенства $\|\widetilde{\nabla} f(x) - \nabla f(x)\|_2^2 \leqslant \varepsilon^2/10$ используется его ослабленный вариант:

$$E_\xi\left[\|\widetilde{\nabla} f(x) - \nabla f(x)\|_2^2\right] \leqslant \frac{\varepsilon^2}{10}.$$

Использование неравенств концентрации меры (см. приложение) позволяет проработать этот момент аккуратнее. На самом деле приведённые ниже результаты (о сходимости в среднем) можно получить и без предположений, которые требуются для возможности применения данных неравенств (субгауссовские хвосты), см. [224]. ♦

2. **(Stochastic Gradient Descent (SGD).)** Покажите, что если вместо градиента $\nabla f(x)$ доступен стохастический градиент $\nabla_x f(x,\xi)$:

$$E_\xi[\nabla_x f(x,\xi)] = \nabla f(x), \quad E_\xi\left[\|\nabla_x f(x,\xi) - \nabla f(x)\|_2^2\right] \leqslant D,$$

то для *пробатченного градиента*

$$\widetilde{\nabla} f(x) = \frac{1}{r}\sum_{l=1}^{r} \nabla_x f(x,\xi^l),$$

где случайные величины $\{\xi^l\}_{l=1}^r$ независимы и одинаково распределены, так же как ξ, имеет место оценка

$$E_{\{\xi^l\}_{l=1}^r}\left[\|\widetilde{\nabla} f(x) - \nabla f(x)\|_2^2\right] \leqslant \frac{D}{r}.$$

Выбирая $h_k \equiv 1/(2L)$ и $r \approx 10D/\varepsilon^2$, с помощью установленной оценки покажите, что не более чем за

$$N = \mathrm{O}\left(\frac{L\cdot(f(x^0) - f(x_*))}{\varepsilon^2}\right)$$

итераций метод сможет найти такую точку x^k, что $E[\|\nabla f(x^k)\|_2^2] \leqslant \varepsilon^2$.

3. **(Метод редукции дисперсии и SPIDER [277, 431, 612].)** Предположим дополнительно, что

$$f(x) = \frac{1}{m}\sum_{l=1}^{m} f_l(x),$$

где все функции $f_l(x)$ имеют L-липшицев градиент в 2-норме. Введём стохастический градиент (здесь все ξ^k, ξ^{k-1}, ξ^{k_j}, … независимы и одинаково равномерно распределены среди чисел $1,\ldots,m$)

$$\nabla_x f(x^k, \{\xi^k\}) = \begin{cases} \dfrac{1}{r}\displaystyle\sum_{j=1}^{r} \nabla f_{\xi^{k_j}}(x^k), & \text{если } r < m \text{ и } k \text{ делится на } q,\\ \nabla f(x^k), & \text{если } m \leqslant r \text{ и } k \text{ делится на } q,\\ \nabla f_{\xi^k}(x^k) - \nabla f_{\xi^k}(x^{k-1}) + \nabla_x f(x^{k-1}, \{\xi^{k-1}\}) & \text{иначе.}\end{cases}$$

где $r \approx 20D/\varepsilon^2$; $q = r$, если $r < m$, и $q = m$, если $m \leqslant r$. Пусть

$$h_k = \begin{cases} \dfrac{\varepsilon^2}{20L\sqrt{D}\|\nabla_x f(x^k, \{\xi^k\})\|_2}, & \text{если } r < m, \\ \dfrac{\varepsilon}{L\sqrt{10m}\|\nabla_x f(x^k, \{\xi^k\})\|_2}, & \text{если } m \leqslant r. \end{cases}$$

Докажите, что для любого k до момента остановки выполнено неравенство

$$E_{\{\xi^k\}}\left[\|\nabla_x f(x^k, \{\xi^k\}) - \nabla f(x)\|_2^2\right] \leqslant \frac{\varepsilon^2}{10}.$$

Получите отсюда, что не более чем за

$$N = \mathrm{O}\left(\min\left\{\frac{L\cdot(f(x^0)-f(x_*))\sqrt{D}}{\varepsilon^3} + \frac{D}{\varepsilon^2}, \frac{L\cdot(f(x^0)-f(x_*))\sqrt{m}}{\varepsilon^2} + m\right\}\right) = \\ = L\cdot(f(x^0)-f(x_*))\mathrm{O}\left(\min\left\{\frac{\sqrt{D}}{\varepsilon^3}, \frac{\sqrt{m}}{\varepsilon^2}\right\}\right)$$

итераций (вычислений $\nabla f_\xi(x)$) метод сможет найти такую точку x^k, что $E[\|\nabla f(x^k)\|_2] \leqslant \varepsilon$.

Указания. 1. Из неравенства (1.5) и сноски 12, в которой выбираем

$$L := \frac{1}{2h_k},$$

следует, что

$$f(x^{k+1}) \leqslant f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k\rangle + \frac{L}{2}\|x^{k+1} - x^k\|_2^2 \leqslant \\ \leqslant f(x^k) + \langle \widetilde{\nabla} f(x^k), x^{k+1} - x^k\rangle + \left(\frac{L}{2} + \frac{1}{4h_k}\right)\|x^{k+1} - x^k\|_2^2 + \\ + h_k\|\widetilde{\nabla} f(x^k) - \nabla f(x^k)\|_2^2 \leqslant f(x^k) - \frac{h_k}{2}\|\widetilde{\nabla} f(x^k)\|_2^2 + \frac{h_k}{10}\varepsilon^2.$$

Если

$$\|\widetilde{\nabla} f(x^k)\|_2^2 > \frac{2\varepsilon^2}{5}, \quad \varepsilon^2\varsigma \leqslant \|\widetilde{\nabla} f(x^k)\|_2^2 h_k,$$

то из последнего неравенства получаем, что

$$f(x^{k+1}) < f(x^k) - \frac{h_k}{2}\|\widetilde{\nabla} f(x^k)\|_2^2 + \frac{h_k}{4}\|\widetilde{\nabla} f(x^k)\|_2^2 \leqslant f(x^k) - \frac{\varsigma}{4}\varepsilon^2.$$

2. (См. начало приложения и [138, 295].) Отметим, что оценка на общее число вычислений $\nabla_x f(x, \xi)$ при сделанных предположениях не может быть улучшена (с точностью до числового множителя) никаким другим методом [138].

3. (См. начало приложения и [277, 620].) Описанный способ формирования стохастического градиента (и близкие к нему способы, см. приложение) в оптимизационном сообществе принято называть

методом редукции дисперсии [282, section 7]. Смысл данного названия поясняет следующее наблюдение (SPIDER [277]): для

$$k = a\cdot q + p, \quad 0 \leqslant p < q \quad \text{и} \quad \|\nabla_x f(x^k, \{\xi^k\})\|_2 h_k \leqslant \tilde{h}$$

справедливы неравенства

$$\begin{aligned} &E_{\{\xi^k\}}[\|\nabla_x f(x^k, \{\xi^k\}) - \nabla f(x)\|_2^2] \leqslant \\ &\quad\leqslant \begin{cases} E_{\{\xi^{a\cdot q}\}}\left[\left\|\nabla_x f(x^{a\cdot q}, \{\xi^{a\cdot q}\}) - \nabla f(x)\right\|_2^2\right] + L^2\tilde{h}^2 p, & \text{если } r < m, \\ L^2\tilde{h}^2 p, & \text{если } m \leqslant r, \end{cases} \leqslant \\ &\quad\leqslant \begin{cases} \dfrac{\varepsilon^2}{20} + \dfrac{\varepsilon^4 p}{400D}, & \text{если } r < m, \\ \dfrac{\varepsilon^4 p}{10m}, & \text{если } m \leqslant r, \end{cases} \leqslant \frac{\varepsilon^2}{10}. \end{aligned}$$

Таким образом, делая q итераций, мы вычисляем $\nabla f_\xi(x)$ всего $3q$ раз. При этом дисперсию стохастического градиента удаётся уменьшить в $\sim D/\varepsilon^2$ раз. Однако за это есть и «плата» в виде «маленьких» шагов, что приводит к увеличению необходимого числа итераций. Тем не менее в итоге получается выиграть на числе вычислений $\nabla f_\xi(x)$ по сравнению с подходом, изложенным в предыдущем пункте.

Отметим, что приведённая оценка на число вычислений $\nabla f_\xi(x)$ при сделанных предположениях не может быть улучшена (с точностью до числового множителя) никаким другим методом [138]. Более того, для полученной оценки (в варианте $r < m$) не требуется выполнения равенства

$$f(x) = \frac{1}{m}\sum_{l=1}^{m} f_l(x).$$

Достаточно, чтобы выполнялось неравенство [138, 277]

$$E_\xi\left[\left\|\nabla_x f(y,\xi) - \nabla_x f(x,\xi)\right\|_2^2\right] \leqslant L^2\|y-x\|_2^2. \tag{1.47}$$

При этом в варианте $m \leqslant r$ не требуется выполнения условия $D < \infty$.

Отметим также, что в предположении ограниченности старших производных оптимизируемой функции (достаточно липшицевости гессиана и возможности умножения оценки гессиана на вектор; привлечение (оценок) старших производных ничего не даёт) можно получить оценку $N \sim \varepsilon^{-3}$ без предположения (1.47) [140].

§ 2

Метод проекции градиента

Рассмотрим задачу выпуклой оптимизации

$$f(x) \to \min_{x \in Q}. \tag{2.1}$$

Это значит, что $f(x)$ — выпуклая функция, а $Q \subseteq \mathbb{R}^n$ — выпуклое множество, которое мы считаем достаточно *простым* в том смысле, что решение вспомогательной задачи проектирования на это множество (1.31) занимает существенно меньше времени, чем расчёт градиента $\nabla f(x)$. В качестве наглядного примера можно рассмотреть задачу минимизации квадратичной функции на параллелепипеде, задав, например,

$$f(x) = \frac{1}{2}\|Ax\|_2^2, \quad Q = \prod_{k=1}^{n} [a_k, b_k].$$

В случае плотной матрицы A расчёт $\nabla f(x) = A^T \cdot (Ax)$ будет стоить $\mathrm{O}(n^2)$ *арифметических операций*[1], а проектирование на Q согласно соотношению (1.31) делается по явным формулам за время $\mathrm{O}(n)$. Совсем необязательно, чтобы проектирование осуществлялось по явным формулам. Однако в подавляющем большинстве рассматриваемых в приложениях случаев простых множеств Q проектирование может быть осуществлено за время (см. упражнение 4.6)

$$\mathrm{O}\left(n \ln^2\left(\frac{n}{\varepsilon}\right)\right), \tag{2.2}$$

где ε — относительная точность проектирования (в смысле сходимости по аргументу или по функции — в данном случае неважно). В противном случае множество уже, как правило, не считают простым, и его стараются описывать с помощью функциональных ограничений.

[1] Операций типа сложения, умножения, деления двух чисел типа *float* [46, п. 1.3] — все эти операции сопоставимы (с точностью до логарифмического множителя от длины операндов) по сложности [63, гл. 29], [67]. Число арифметических операций определяет время работы программы (длительность вычислений, трудоёмкость), поэтому далее вместо числа арифметических операций также будет использоваться словосочетание *время работы*.

Тогда становится правильнее говорить уже о задаче *условной оптимизации* [86], см. также пример 3.2, замечание 4.3 и упражнение 5.5.

Существенным недостатком подхода из § 1 является предположение о том, что неравенство (1.4) имеет место на всём пространстве $\mathbb{R}^n$. Легко понять, что это довольно обременительное условие. Например, простая выпуклая функция скалярного аргумента $f(x) = x^4$ не удовлетворяет этому условию. Далее в этом параграфе путём специальной компактификации, вообще говоря, неограниченного множества Q мы избавимся от отмеченной проблемы.

Другим недостатком подхода из § 1 является невозможность его использования для выпуклых, но негладких функций $f(x)$. Действительно, следуя [86, § 3, гл. 5], рассмотрим

$$f(x_1, x_2) = |x_1 - x_2| + 0{,}2|x_1 + x_2|.$$

Естественно пытаться заменять градиент в подходе из § 1 субградиентом (произвольным элементом субдифференциала) в точках, в которых $f(x)$ не является гладкой.

♦ Напомним, что *субдифференциал* — это в общем случае выпуклый компакт [66, п. 1.5]. Например, для функции скалярного аргумента $f(x) = |x|$ субдифференциал будет иметь вид

$$\partial f(x) = -1,\ x < 0; \quad \partial f(x) = 1,\ x > 0; \quad \partial f(x) = [-1, 1],\ x = 0.$$

Напомним также, что мера (Лебега) точек негладкости выпуклой функции равна нулю по теореме Радемахера [105, 452], однако часто решение негладких задач достигается в одной из таких точек, и получается, что градиентный спуск может проводить заметную долю времени в окрестности таких точек. ♦

Рассмотрим одну из таких точек $(1, 1)$. Используя, например, соотношение (1.17), несложно проверить, что вектор (субградиент)

$$\nabla f(1, 1) = (1{,}2; -0{,}8)$$

будет принадлежать субдифференциалу $\partial f(1, 1)$. Однако при любом выборе шага $h > 0$ в методе (1.3) функция $f(x)$ из точки $(1, 1)$ может только возрастать по направлению $\nabla f(1, 1)$. Таким образом, в негладком случае рассчитывать на основное неравенство (1.7) не приходится, что и неудивительно, поскольку в это неравенство входит константа Липшица градиента L, предполагающая гладкость функции $f(x)$.

Более того, рассматривая простейшую негладкую выпуклую функцию скалярного аргумента с острым минимумом $f(x) = |x|$, получаем,

что $|\partial f(x)| = 1$, если только случайно мы не оказались в точке $x = 0$. Поэтому для метода (1.3) с шагом h для почти всех точек старта x^0 имеем $|x^{k+1} - x^k| = h$, откуда следует, что для любого k выполняется неравенство

$$\max\{f(x^k), f(x^{k+1})\} \geqslant \frac{h}{2}.$$

Значит, необходимо выбирать h пропорционально желаемой точности решения задачи ε либо считать, что $h_k \to 0$ при $k \to \infty$, чтобы оказаться в нужной окрестности решения. Это существенно отличается от способа выбора шага (1.6) в гладком случае.

Несмотря на отмеченные сложности, далее, следуя Ю. Е. Нестерову [492], мы постараемся единообразно посмотреть на гладкий и негладкий случаи.

Определим множество $B_{R,Q}(x_*) = \{x \in Q : \|x - x_*\|_2 \leqslant R\}$, где x_* — решение задачи (2.1), $R = \|x^0 - x_*\|_2$. Если решение не единственно, то под x_* будем понимать такое решение задачи (2.1), которое наиболее близко в 2-норме к точке старта x^0. Предположим, что для любых $x, y \in B_{R,Q}(x_*)$ выполнено неравенство

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2 + \delta, \tag{2.3}$$

где $\delta > 0$. В частности, если градиент функции $f(x)$ удовлетворяет *условию Гёльдера*, точнее, для любых $x, y \in B_{R,Q}(x_*)$ имеет место неравенство

$$\|\nabla f(y) - \nabla f(x)\|_2 \leqslant L_\nu \|y - x\|_2^\nu, \quad \nu \in [0, 1], \quad L_0 < \infty, \tag{2.4}$$

то неравенство (2.3) имеет место с

$$L = L_\nu \cdot \left[\frac{L_\nu}{2\delta}\frac{1-\nu}{1+\nu}\right]^{\frac{1-\nu}{1+\nu}}. \tag{2.5}$$

Детали см. в работе [235]. См. также рис. 7, соответствующий $\nu = 0$.

♦ В случае $\nu = 0$ условие (2.4) (аналогично (2.3)) выполняется для любых элементов соответствующих субдифференциалов $\partial f(x)$ и $\partial f(y)$. Фактически условие (2.4) отвечает тому, что у функции $f(x)$ равномерно ограниченны все элементы субдифференциалов во всех точках, т. е. она имеет равномерно ограниченную константу Липшица. Заметим, что с небольшими оговорками это условие отвечает самому общему классу всех собственных выпуклых функций на открытом множестве, содержащем $B_{R,Q}(x_*)$, см. также [158, 319].

Отметим также, что с точки зрения формальной логики формула (2.4) некорректна, потому что она незамкнута относительно параметра ν [83]. Однако это было сделано вполне осмысленно. Дело в том,

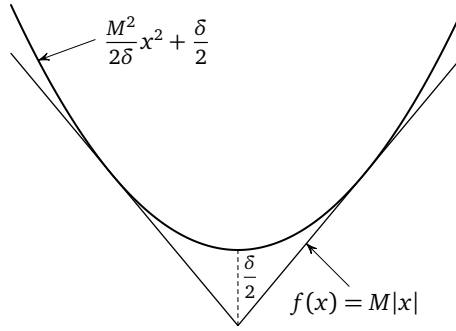

Рис. 7

что в настоящем параграфе на формулу (2.4) мы будем смотреть только с точки зрения конкретного ν. А в § 5 мы уже будем «играть» на выборе параметра $\nu\in[0,1]$, считая, что неравенство (2.4) имеет место для любого $\nu\in[0,1]$, но при этом допуская возможность того, что $L_\nu=\infty$ начиная с некоторого $\nu\in(0,1]$.

Во всех последующих формулах, если не оговорено противное (см. соотношения (3.4), (3.5)), использование без уточнений $\nabla f(x)$ подразумевает, что формулы справедливы при любом выборе $\nabla f(x)\in$ $\in\partial f(x)$. ♦

Рассмотрим (см. также формулу (1.31)) простейший метод проекции градиента с шагом $h\leqslant 1/L$:

$$x^{k+1}=\pi_Q(x^k-h\nabla f(x^k))=\arg\min_{x\in Q}\left\{\langle h\nabla f(x^k),x-x^k\rangle+\frac{1}{2}\|x-x^k\|_2^2\right\}=$$
$$=\arg\min_{x\in Q}\left\{f(x^k)+\langle\nabla f(x^k),x-x^k\rangle+\frac{1}{2h}\|x-x^k\|_2^2\right\}. \quad (2.6)$$

Следствием соотношения (2.6) является условие, которое получается из неравенства (1.17) при $x=x^{k+1}$, $y=x$: для всех $x\in Q$ выполнено соотношение

$$\left\langle\nabla_x\left(\langle h\nabla f(x^k),x-x^k\rangle+\frac{1}{2}\|x-x^k\|_2^2\right)\Big|_{x=x^{k+1}},x-x^{k+1}\right\rangle=$$
$$=\left\langle h\nabla f(x^k)+x^{k+1}-x^k,x-x^{k+1}\right\rangle\geqslant 0, \quad (2.7)$$

т. е.

$$\langle h\nabla f(x^k),x^{k+1}-x\rangle\leqslant\langle x^{k+1}-x^k,x-x^{k+1}\rangle=$$
$$=\frac{1}{2}\|x-x^k\|_2^2-\frac{1}{2}\|x-x^{k+1}\|_2^2-\frac{1}{2}\|x^{k+1}-x^k\|_2^2. \quad (2.8)$$

Следуя работе [127], введём

$$\mathrm{Prog}^h(x^k) \stackrel{\text{def}}{=} -\min_{x\in Q}\left\{\langle \nabla f(x^k), x-x^k\rangle + \frac{1}{2h}\|x-x^k\|_2^2\right\} \geqslant 0. \tag{2.9}$$

По формуле (2.6) для всех $x\in Q$ имеет место следующее «правильное» обобщение равенства (1.18) (следует сравнить с «неправильным»/«грубым» вариантом (1.32)):

$$\begin{aligned}
&\langle h\nabla f(x^k), x^k - x\rangle = \langle h\nabla f(x^k), x^k - x^{k+1}\rangle + \langle h\nabla f(x^k), x^{k+1} - x\rangle \leqslant\\
&\leqslant \langle h\nabla f(x^k), x^k - x^{k+1}\rangle + \frac{1}{2}\|x-x^k\|_2^2 - \frac{1}{2}\|x-x^{k+1}\|_2^2 - \frac{1}{2}\|x^{k+1}-x^k\|_2^2 =\\
&= -h\left\{\langle \nabla f(x^k), x^{k+1}-x^k\rangle + \frac{1}{2h}\|x^{k+1}-x^k\|_2^2\right\} +\\
&\qquad\qquad + \frac{1}{2}\|x-x^k\|_2^2 - \frac{1}{2}\|x-x^{k+1}\|_2^2 \leqslant\\
&\leqslant h\,\mathrm{Prog}^h(x^k) + \frac{1}{2}\|x-x^k\|_2^2 - \frac{1}{2}\|x-x^{k+1}\|_2^2.
\end{aligned} \tag{2.10}$$

Если $x^k, x^{k+1}\in B_{R,Q}(x_*)$, то с учётом условия $h\leqslant 1/L$ из соотношения (2.3) получаем

$$\begin{aligned}
\mathrm{Prog}^h(x^k) &= -\min_{x\in Q}\left\{\langle \nabla f(x^k), x-x^k\rangle + \frac{1}{2h}\|x-x^k\|_2^2\right\} =\\
&= -\left(\langle \nabla f(x^k), x^{k+1}-x^k\rangle + \frac{1}{2h}\|x^{k+1}-x^k\|_2^2\right) =\\
&= f(x^k) - \underbrace{\left(f(x^k) + \langle \nabla f(x^k), x^{k+1}-x^k\rangle + \frac{1}{2h}\|x^{k+1}-x^k\|_2^2 + \delta\right)}_{\geqslant f(x^{k+1})} + \delta \leqslant\\
&\leqslant f(x^k) - f(x^{k+1}) + \delta.
\end{aligned} \tag{2.11}$$

Подставляя соотношение (2.11) в неравенство (2.10), в предположении, что $x^k, x^{k+1}\in B_{R,Q}(x_*)$, аналогично неравенству (1.19) получим

$$h\langle \nabla f(x^k), x^k - x\rangle \leqslant h\cdot(f(x^k) - f(x^{k+1}) + \delta) + \frac{1}{2}\|x-x^k\|_2^2 - \frac{1}{2}\|x-x^{k+1}\|_2^2. \tag{2.12}$$

В силу выпуклости функции $f(x)$ имеем (см. формулу (1.17))

$$f(x^k) - f(x) \leqslant \langle \nabla f(x^k), x^k - x\rangle, \tag{2.13}$$

а также

$$f(\bar{x}^m) \leqslant \frac{1}{m}\sum_{k=1}^m f(x^k), \tag{2.14}$$

где (см. формулу (1.21))

$$\bar{x}^m = \frac{1}{m}\sum_{k=1}^m x^k.$$

В неравенстве (2.12) положим $x = x_*$, а если решение не единственно, то выберем то x_*, для которого $\|x^0 - x_*\|_2^2$ минимально.

Суммируя неравенства (2.12), записанные с учётом оценки (2.13):

$$h\cdot(f(x^k)-f(x_*)) \leqslant h\cdot(f(x^k)-f(x^{k+1})+\delta)+\frac{1}{2}\|x_*-x^k\|_2^2-\frac{1}{2}\|x_*-x^{k+1}\|_2^2,$$

т. е.

$$h\cdot(f(x^{k+1})-f(x_*)) \leqslant h\delta+\frac{1}{2}\|x_*-x^k\|_2^2-\frac{1}{2}\|x_*-x^{k+1}\|_2^2, \tag{2.15}$$

по $k = 0, \ldots, m-1$, с учётом неравенства (2.14) получим

$$mh\cdot(f(\bar{x}^m)-f(x_*)) \leqslant mh\delta+\frac{1}{2}\|x_*-x^0\|_2^2-\frac{1}{2}\|x_*-x^m\|_2^2, \tag{2.16}$$

т. е.

$$\frac{1}{2}\|x_*-x^m\|_2^2 \leqslant mh\cdot\left(\delta-(f(\bar{x}^m)-f(x_*)\right)+\frac{1}{2}\|x_*-x^0\|_2^2. \tag{2.17}$$

Вполне естественно (см. § 4, 5) рассчитывать на то, что метод останавливается, когда

$$\varepsilon = f(\bar{x}^N)-f(x_*) \approx 2\delta > \delta, \tag{2.18}$$

где

$$\bar{x}^N = \frac{1}{N}\sum_{k=1}^{N} x^k.$$

Как мы увидим далее (см. неравенство (2.22) и упражнение 2.2), получить точность $\varepsilon < \delta$ в общем случае не представляется возможным. Поэтому из оценок (2.17) и (2.18) имеем

$$\frac{1}{2}\|x_*-x^k\|_2^2 \leqslant \frac{1}{2}\|x_*-x^0\|_2^2, \quad k = 0, \ldots, N. \tag{2.19}$$

Другими словами, если $x^0 \in B_{R,Q}(x_*)$ (а это выполняется по построению $B_{R,Q}(x_*)$), то для любого $k = 0, \ldots, N$ также верно, что $x^k \in B_{R,Q}(x_*)$. Таким образом, оговорку о том, что $x^k, x^{k+1} \in B_{R,Q}(x_*)$, можно опустить.

Строго говоря, мы вывели этот факт, как бы на него же и опираясь (см. оговорку «в предположении, что $x^k, x^{k+1} \in B_{R,Q}(x_*)$» около формулы (2.12)). Однако несложно понять, что, предполагая условие (2.3) выполненным и вне множества $B_{R,Q}(x_*)$, т. е. на всём Q, с теми же параметрами (L, δ), мы уже без всяких оговорок получаем, что $x^k \in B_{R,Q}(x_*)$. Но это означает, что последовательность $\{x^k\}_{k=0}^N$ никогда не выйдет за пределы множества $B_{R,Q}(x_*)$, и поэтому от того, что именно мы предполагали о выпуклой на всём множестве Q функции $f(x)$ вне множества $B_{R,Q}(x_*)$, ничего не зависит — мы никогда не окажемся вне $B_{R,Q}(x_*)$.

Вернёмся к формуле (2.16) при $m = N$, которую перепишем следующим образом:

$$f(\bar{x}^N) - f(x_*) \leqslant \frac{1}{2hN}\|x_* - x^0\|_2^2 + \delta. \tag{2.20}$$

Вспоминая, что h должно было удовлетворять условию $h \leqslant 1/L$, аналогично (1.6) выберем

$$h = \frac{1}{L}. \tag{2.21}$$

Подставляя выражение (2.21) в неравенство (2.20), аналогично (1.20) получим

$$f(\bar{x}^N) - f(x_*) \leqslant \frac{LR^2}{2N} + \delta. \tag{2.22}$$

Замечание 2.1 (условие слабой квазивыпуклости). Вместо $\bar{x}^N$ в приведённых выше формулах можно использовать

$$\widehat{x}^N = \arg\min_{k=1,\ldots,N} f(x^k). \tag{2.23}$$

Несложно заметить, что в случае подхода с $\widehat{x}^N$ приведённые выше рассуждения используют лишь свойство (2.13) с $x = x_*$:

$$(f(x^k) - f(x_*)) \leqslant \langle \nabla f(x^k), x^k - x_* \rangle,$$

т. е. «полноценная» выпуклость функции $f(x)$ не требуется. По-видимому, это наблюдение восходит к Н. З. Шору [86, п. 5.4.4].

Отметим, что условие (2.13) также ослабляют следующим образом (следует сравнить с однородными относительно x_* функциями [86, п. 4, § 3, гл. 3] и звёздной выпуклостью [345, 480]):

$$\alpha \cdot (f(x^k) - f(x_*)) \leqslant \langle \nabla f(x^k), x^k - x_* \rangle, \quad \alpha \in (0, 1]. \tag{2.24}$$

Условие (2.24) иногда называют *условием α-слабой квазивыпуклости* функции $f(x)$. В последнее время оно стало достаточно популярно в связи с приложениями, возникающими в *глубоком обучении* (*Deep Learning*) [331]. Несложно показать, что приведённые выше рассуждения переносятся и на этот случай. При этом оценка (2.22) «портится» следующим образом [324]:

$$f(\widehat{x}^N) - f(x_*) \leqslant \frac{LR^2}{2\alpha N} + \delta,$$

где $\widehat{x}^N$ определяется формулой (2.23).

В приложениях часто используют также такое неравенство:

$$f(\bar{x}^N) - f(x_*) \leqslant \underbrace{\sup_{x\in Q} \frac{1}{N}\sum_{k=0}^{N-1}\langle \nabla f(x^k), x^k - x\rangle}_{\text{сертификат точности}} \leqslant$$

$$\leqslant \frac{f(x^0) - f(x^N) + L\sup_{x\in Q}\|x - x^0\|_2^2}{2N} + \delta, \qquad (2.25)$$

где (следует сопоставить с формулой (1.21))

$$\bar{x}^N = \frac{1}{N}\sum_{k=0}^{N-1} x^k.$$

Введённый в оценке (2.25) *сертификат точности* (accuracy certificate) играет ключевую роль в обосновании прямодвойственности исследуемого метода [465] (см. также § 4). Сертификат точности и его аналоги из § 4 являются вычислимыми (не требуют знания, как правило, неизвестных значений $f(x_*)$ или R^2) и потому могут использоваться в качестве критерия останова методов. ■

Предположим, что неравенство (2.3) имеет вид (см. также замечание 1.3)

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2 + \delta. \qquad (2.26)$$

Для этого, например, достаточно, чтобы имело место неравенство (см. также (1.26)), аналогичное (2.4):

$$\|\nabla f(y) - \nabla f(x)\|_* \leqslant L_\nu \|y - x\|^\nu, \quad \nu \in [0,1],\ L_0 < \infty. \qquad (2.27)$$

Тогда неравенство (2.26) имеет место с константой L, рассчитываемой по формуле (2.5). Попробуем, следуя А. С. Немировскому [164], распространить метод градиентного спуска на этот случай. Как уже отмечалось в предыдущем параграфе, сделать это за счёт такого обобщения метода не получается (см. формулу (1.27)):

$$x^{k+1} = \arg\min_{x\in Q}\left\{\langle h\nabla f(x^k), x - x^k\rangle + \frac{1}{2}\|x - x^k\|^2\right\}. \qquad (2.28)$$

Причина прежде всего в том, что, например, $\|x - x^k\|_1^2$ не есть сильно выпуклая функция в 2-норме и тем более в 1-норме. Отсутствие этого свойства, как мы увидим чуть ниже, и не позволяет сделать необходимое обобщение. Рассмотрим, однако, близкий к (2.28) метод

$$x^{k+1} = \arg\min_{x\in Q}\left\{\langle h\nabla f(x^k), x - x^k\rangle + V(x, x^k)\right\}. \qquad (2.29)$$

Получим условия на выпуклую по x функцию $V(x, x^k)$, при которых приведённая в начале параграфа *конструкция* вывода основных оценок *сохраняется*. Первым ключевым местом, в котором использовались свойства функции $V(x, y) = \|x - y\|_2^2/2$, было неравенство (2.8). В случае (2.29) неравенство (2.8) должно было бы принять вид

$$\langle h\nabla f(x^k), x^{k+1} - x\rangle \leqslant \left\langle \nabla_{x^{k+1}} V(x^{k+1}, x^k), x - x^{k+1}\right\rangle \overset{①}{=} \\ \overset{①}{=} V(x, x^k) - V(x, x^{k+1}) - V(x^{k+1}, x^k). \qquad (2.30)$$

Таким образом, достаточно потребовать выполнение равенства[2] ① тождественно по x. Например, если считать, что имеет место представление

$$V(x, y) = d(x) - d(y) - \langle \nabla d(y), x - y\rangle \qquad (2.31)$$

с выпуклой функцией $d(x)$, то тождество ① также имеет место. Вторым и заключительным ключевым местом было неравенство (2.12), которое в нашем случае останется верным, если

$$V(x^{k+1}, x^k) \geqslant \frac{1}{2}\|x^{k+1} - x^k\|^2. \qquad (2.32)$$

Для этого достаточно, чтобы в представлении (2.31) функция $d(x)$ была 1-сильно выпукла в выбранной норме $\|\ \|$. Функцию $d(x)$ называют *прокс-функцией*, а функцию $V(x, y)$ — порождённым ею *расхождением* или *дивергенцией Брэгмана* (Bregman divergence) [12, 164, 186]. Отметим, что для метода (2.30) в оценку (2.23) будет входить $R^2 = 2V(x_*, x^0)$. Если решение не единственно, то оценка (2.23) будет верна в том числе и для того решения x_*, которое доставляет минимум R^2. Отметим также, что для «сохранения конструкции» достаточно, чтобы условия (2.27), (2.28) имели место только при

$$x, y \in \{x \in Q : V(x_*, x) \leqslant R^2\}.$$

Рассуждения здесь аналогичны рассуждениям, использованным при выводе оценки (2.20).

Примеры прокс-функций для множеств Q вида шаров в различных нормах собраны в табл. 1 [20, 164]. Параметр a имеет вид

$$a = \frac{2\ln n}{2\ln n - 1} \simeq 1 + \frac{1}{2\ln n}.$$

[2] Условие на функцию $V(x, y)$, накладываемое равенством 1, можно понимать и как неравенство в сторону «$\leqslant$». Однако, как будет видно в дальнейшем, удаётся подобрать функцию $V(x, y)$ и с равенством, что дополнительно привносит меньше грубости в рассуждения.

Таблица 1

$Q = B_p^n(1)$	$1 \leqslant p \leqslant a$	$a \leqslant p \leqslant 2$	$2 \leqslant p \leqslant \infty$
$\|\ \|$	$\|\ \|_1$	$\|\ \|_p$	$\|\ \|_2$
$d(x)$	$\frac{1}{2(a-1)}\|x\|_a^2$	$\frac{1}{2(p-1)}\|x\|_p^2$	$\frac{1}{2}\|x\|_2^2$
R^2	$O(\ln n)$	$O((p-1)^{-1})$	$O(n^{1/2-1/p})$

Дополнительно к тому, что приведено в табл. 1, особо отметим «Spectrahedron setup» (спектраэдральную прокс-структуру) [186, п. 4.3]. Приведённые в табл. 1 прокс-функции можно распространить и на прямые произведения шаров [164, п. 5.3.3].

По-видимому, в табл. 1 в общем случае нельзя избавиться от дополнительного фактора (множителя) $\ln n$ в оценке R^2 с помощью дивергенции Брэгмана по сравнению с оценкой R^2, равной квадрату соответствующей нормы. Однако эта плата позволяет переносить все основные свойства, присущие евклидову случаю, на множества типа шаров в 1-норме. Кроме того, во всех использующихся примерах прокс-структур эта мультипликативная плата по порядку не превышает $\ln n$. Такой порядок у этой константы будет, например, для единичного симплекса (см. также текст ниже) при

$$d(x) = \sum_{i=1}^n x_i \ln x_i, \quad V(x,y) = \sum_{i=1}^n x_i \ln\Big(\frac{x_i}{y_i}\Big) \quad \text{и} \quad x^0 = (n^{-1}, \dots, n^{-1}).$$

Все приведённые в табл. 1 прокс-функции 1-сильно выпуклы в указанных нормах на всём пространстве. Поэтому их можно использовать и в том случае, когда мы заранее не знаем, где локализовано решение [3]. Например, если стартовать с точки $x^0 = 0$ и заранее знать, что решение разреженно (большинство компонент равно нулю), то естественно выбирать 1-норму и соответствующую прокс-функцию (см. табл. 1). Действительно, в этом случае можно ожидать, что R^2 в оценке (2.23) не будет сильно зависеть от выбора нормы, в то время как для константы

$$L := L^p = \sup_{x \in Q} \max_{\|h\|_p \leqslant 1} \langle h, \nabla^2 f(x) h \rangle$$

отличие может быть в n раз, поскольку $L^2/n \leqslant L^1 \leqslant L^2$. Скажем, для функции (1.30)

$$f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$$

несложно получить, что $L^1 = \max_{i,j=1,\dots,n} |A_{ij}|$, а $L^2 = \lambda_{\max}(A)$.

♦ Аналогично можно определить и константу сильной выпуклости относительно p-нормы

$$\mu^p = \inf_{x\in Q} \min_{\|h\|_p\leqslant 1} \langle h, \nabla^2 f(x)h\rangle.$$

Строго говоря, приведённые здесь определения констант L^p и μ^p справедливы только при дополнительном предположении о существовании и конечности матрицы Гессе $\nabla^2 f(x)$ при $x\in Q$. Последнее предположение не всегда имеет место. В частности, функция Хьюбера (1.43) имеет липшицев градиент, однако гессиан определён не везде. ♦

Как мы увидим в дальнейшем (см. упражнение 4.6), задача (2.30) при известном векторе $\nabla f(x^k)$ для примеров из табл. 1 решается за время $O(n\ln^2(n/\varepsilon))$, где ε — точность решения в смысле сходимости по аргументу. Выделим особо один частный случай, когда эту оценку можно улучшить до $O(n)$:

$$Q = S_n(1) = \left\{x\in\mathbb{R}^n_+ \colon \sum_{i=1}^n x_i = 1\right\},$$

т. е. Q — единичный симплекс в $\mathbb{R}^n$. В этом случае, выбирая

$$d(x) = \sum_{i=1}^n x_i \ln x_i, \tag{2.33}$$

получим

$$x_i^{k+1} = \frac{x_i^k \exp\left(-h\dfrac{\partial f(x^k)}{\partial x_i}\right)}{\displaystyle\sum_{j=1}^n x_j^k \exp\left(-h\frac{\partial f(x^k)}{\partial x_j}\right)}, \quad i = 1,\ldots,n.$$

Отметим, что метод (2.30) при $h\leqslant 1/L$ ввиду неравенства

$$V(x,y) \geqslant \frac{\|x-y\|^2}{2}$$

имеет геометрическую интерпретацию, аналогичную обычному градиентному спуску (см. замечания 1.2, 1.3).

Конструкция (2.8)–(2.13) с учётом соотношений (2.30), (2.31) может быть распространена на более общий класс задач. В следующем параграфе приводится основная схема такого распространения, лежащая в основе получения результатов в наиболее общем виде.

Упражнение 2.1 (нижние оценки — негладкий случай). Покажите, что в условии (2.5) с $\nu = 0$ оценка (2.23) для метода (2.7), (2.22) с

$$h = \frac{1}{L} = \frac{2\delta}{L_0^2} = \frac{\varepsilon}{L_0^2}, \quad \text{где} \quad \varepsilon = \frac{LR^2}{2N} + \delta = \frac{LR^2}{2N} + \frac{\varepsilon}{2},$$

т. е.[3] с

$$h = \frac{\varepsilon}{L_0^2} = \frac{R}{L_0\sqrt{N}}, \tag{2.34}$$

будет иметь вид

$$f(\bar{x}^N) - f(x_*) \leqslant \frac{L_0 R}{\sqrt{N}}. \tag{2.35}$$

Покажите, что в классе методов

$$x^{k+1} \in x^0 + \operatorname{Lin}\{\partial f(x^0), \ldots, \partial f(x^k)\} \tag{2.36}$$

оценка (2.35) при $N \leqslant n-1$, где $n = \dim x$, не может быть улучшена с точностью до числового множителя[4]. Предполагается, что в рассматриваемом классе методов (2.36) нельзя выбирать субградиент из субдифференциала. Это означает, что при оценивании скорости сходимости необходимо исходить из того, что субградиент из субдифференциала может выбираться наиболее неудачным образом.

Указание. Следует сравнить это упражнение с упражнением 1.3. Согласно соотношению (2.6) имеем $L = L_0^2/(2\delta)$. С учётом этого найдите минимум правой части неравенства (2.23) по $\delta > 0$. Получите отсюда оценку (2.35).

Чтобы получить нижнюю оценку, воспользуйтесь, например, [76, п. 3.2.1], [186, п. 3.5]. По заданному $N \leqslant n-1$ определите

$$f(x) = F_{N+1}(x) = L_0 \max_{1 \leqslant i \leqslant N+1} x_i + \frac{\mu}{2}\|x\|_2^2, \quad \mu = \frac{L_0}{R\sqrt{N+1}}.$$

Из решения задачи

$$L_0\tau + \frac{\mu \cdot (N+1)}{2}\tau^2 \to \min_{\tau}$$

определите

$$\tau_* = -\frac{R}{\sqrt{N+1}}, \quad x_* = (\underbrace{\tau_*, \ldots, \tau_*}_{N+1}, 0, \ldots, 0).$$

Тогда

$$\|x_*\|_2^2 = (N+1)\tau_*^2 = R^2, \quad f(x_*) = \min_{x \in \mathbb{R}^n} F_{N+1}(x) = -\frac{L_0 R}{(2\sqrt{N+1})} = F_{N+1}^*.$$

---

[3] То, что для негладких задач шаг градиентного метода $h = \text{const} \cdot \varepsilon/L_0^2$, а для гладких — $h = \text{const}/L_1$, можно было понять и из П-теоремы теории размерностей [4, 53]. Отметим также, что в негладком случае ещё возможен вариант $h = \text{const} \cdot R/L_0$. Подробнее об этом см. в упражнении 2.6.

[4] На самом деле, если ограничиться только классом методов вида (2.36) с фиксированными шагами (не зависящими от оптимизируемой функции, см. замечание 1.5), в оценке (2.35) можно немного улучшить только знаменатель в правой части неравенства: $\sqrt{N} \to \sqrt{N+1}$, см. [248].

Если $x^0 = 0$, то для метода вида (2.36) после $k \leqslant N$ итераций при специальном (наиболее неблагоприятном) выборе субградиентов из субдифференциалов имеет место условие $x_i^k = 0$ при $i > k$. Действительно, если это условие верно для шага $k-1$, то для того, чтобы появились новые ненулевые компоненты у вектора x^k (на шаге k) по сравнению с x^{k-1}, необходимо, чтобы выполнялось неравенство $\max_{1\leqslant i\leqslant N} x_i^{k-1} \leqslant 0$. В этом случае субдифференциал $\partial \max_{1\leqslant i\leqslant N} x_i^{k-1}$ будет определяться выпуклой комбинацией таких единичных ортов e_i, для которых $x_i^{k-1} = 0$. В частности, можно взять $\partial \max_{1\leqslant i\leqslant N} x_i^{k-1} = e_k$. Такой выбор (в случае, если $\max_{1\leqslant i\leqslant N} x_i^{k-1} < 0$) обусловлен желанием обеспечить как можно более медленную скорость сходимость метода вида (2.36), что в конечном итоге должно приводить к наиболее точным нижним оценкам. Таким образом, поскольку $x_{N+1}^N = 0$, получаем, что $\max_{1\leqslant i\leqslant N+1} x_i^N = 0$. Значит, $F_{N+1}(x^N) \geqslant 0$. Следовательно,

$$F_{N+1}(x^N) - F_{N+1}^* \geqslant -F_{N+1}^* = \frac{L_0 R}{2\sqrt{N+1}} = \frac{L_0^2}{2\mu\cdot(N+1)}. \tag{2.37}$$

Заметим, что оценка (2.37) одновременно является нижней оценкой в классе методов (2.36) для μ-сильно выпуклых задач в 2-норме. Оценка вида (2.37) будет нижней и для более общего класса методов [74, гл. 4].

Стоит сделать несколько замечаний. Исходная задача — задача безусловной оптимизации. Таким образом, под R в неравенстве (2.37) стоит понимать расстояние от точки старта до решения, собственно, как и в неравенстве (2.35). Не существует сильно выпуклой функции, заданной на всём пространстве, у которой была бы равномерно ограничена константа Липшица (см. также конец § 5). Однако константы, которые входят в оценку (2.37), характеризуют функцию в шаре $B_R(x^0)$, в котором ввиду неравенства (2.20) естественно было бы ожидать пребывание всей траектории метода вида (2.36). В действительности константа L_0, входящая в оценку (2.37), немного меньше настоящей константы Липшица функции $F_N(x)$ в шаре $B_R(x^0)$ ввиду наличия у $F_N(x)$ композитного (сильно выпуклого) слагаемого $\mu\|x\|_2^2/2$. Несложно учесть это слагаемое и должным образом скорректировать оценку (2.37). Общий вывод при этом сохранится [76, п. 3.2.1], [186, п. 3.5]. Однако ввиду примера 3.1 полезно заметить, что и в исходном виде оценка (2.37) представляет ценность, поскольку правильно отражает сложность класса задач композитной выпуклой оптимизации.

Упражнение 2.2. Покажите, что при $\delta > 0$ оценку (2.23) нельзя принципиально улучшить в части зависимости от N, не ухудшив в части зависимости её от $\delta > 0$.

Указание. Проведём рассуждения, следуя работе [235]. Допустите противное, т. е. предположите, что можно получить следующую оценку:

$$f(\bar{x}^N) - f(x_*) \leqslant C_1 \frac{LR^2}{N^{1+\gamma}} + C_2\delta, \quad \gamma > 0. \tag{2.38}$$

Рассмотрите негладкий случай, т. е. используйте равенство (2.5) с $\nu = 0$. Согласно нижним оценкам (см. упражнение 2.1) для любого $N \leqslant n$ и для любого метода вида (2.36) существует такая выпуклая функция из гёльдерова класса с $\nu = 0$ и константой L_0, что

$$f(\bar{x}^N) - f(x_*) \geqslant \frac{L_0 R}{2\sqrt{N}}.$$

Покажите, что при достаточно большом N (а следовательно, и n) это противоречит неравенству (2.38). Для этого согласно соотношению (2.6) подставьте в формулу (2.38) значение $L = L_0^2/(2\delta)$ и специально подберите

$$\delta = L_0 R \sqrt{\frac{C_1}{2C_2 N^{1+\gamma}}}.$$

Тогда

$$f(\bar{x}^N) - f(x_*) \leqslant C_1 \frac{L_0^2 R^2}{2\delta N^{1+\gamma}} + C_2\delta = \sqrt{2C_1C_2}\,\frac{L_0 R}{\sqrt{N^{1+\gamma}}}.$$

Упражнение 2.3 (техника рестартов). 1. Как из метода, работающего по оценке, аналогичной (2.35):

$$f(\bar{x}^N) - f(x_*) \leqslant \frac{L_0 \|x_* - x^0\|_2}{\sqrt{N}} + \delta,$$

где $\delta > 0$ достаточно мало, получить метод, который для μ-сильно выпуклых задач в 2-норме работает по оценке

$$f(\tilde{x}^N) - f(x_*) \leqslant \frac{128 L_0^2}{\mu N} + 2\delta,$$

точнее, по оценке

$$f(\tilde{x}^N) - f(x_*) \leqslant \frac{256 L_0^2}{\mu N}, \quad N \leqslant \frac{128 L_0^2}{\mu\delta}? \tag{2.39}$$

♦ При $\delta = 0$ существует много способов уменьшения константы 256 в оценке (2.39) на два порядка [335, 366, 369, 400, 519]. Однако при этом следует отметить, что согласно упражнению 2.2 оценка (2.39) неулучшаема с точностью до мультипликативной константы. Это общее свойство техники рестартов, описанной в указании к этому упражнению: *из оптимального метода для выпуклых задач с помощью рестартов получают оптимальный метод для сильно выпуклых задач*.

Во всяком случае пока не удалось придумать контрпример, равно как не удалось придумать ситуацию с перенесением метода на сильно выпуклые задачи, в которой бы не было своего варианта рестарт-метода. В основе техники рестартов лежит простая идея: рестартовать метод, т. е. запускать по-новому, возможно, с новыми значениями параметров, каждый раз в момент, когда есть гарантия, что генерируемая последовательность оказалась на «расстоянии» (или невязке по функции) в два раза ближе к решению по сравнению с моментом последнего рестарта. Важно подчеркнуть, что основное свойство описываемой техники (в плане её обоснования) существенно завязано именно на рестарты по расстоянию от текущей точки до решения (или невязке по функции). Использование более удобных критериев для рестарта, например использование легко вычислимой величины нормы градиента функционала для гладких выпуклых задач безусловной оптимизации, к сожалению, не позволяет строго обосновать сохранение свойства оптимальности метода в общем случае [496], см. также замечание 5.3. ♦

2. Попробуйте обобщить этот результат на случай, когда используется произвольная норма $\|\ \|$ и неевклидова прокс-структура, однако имеет место следующее свойство: $d(x) \leqslant C_n\|x\|^2$. Покажите, что при таком предположении оценка (2.39) немного ухудшится: $\mu \to \mu/(2C_n)$. Заметим, что для прокс-функций из табл. 1 имеет место оценка $C_n = \mathrm{O}(\ln n)$. (По-видимому, такой оценки C_n можно добиться во всех интересных для практики случаях при правильном выборе прокс-функции. Заметим, что выбор прокс-функции (2.33) для $Q = S_n(1)$ в этом смысле не будет правильным.)

Указание. 1. См., например, [71, 74, 366, 369, 462, 471, 532]. Покажите, что при $N_1 \leqslant L_0^2R_0^2/\delta^2$ выполняется неравенство

$$\frac{\mu}{2}\underbrace{\|\bar{x}^{N_1}-x_*\|_2^2}_{R_1^2} \overset{①}{\leqslant} f(\bar{x}^{N_1})-f(x_*) \overset{②}{\leqslant} \frac{L_0\overbrace{\|x^0-x_*\|_2}^{R_0}}{\sqrt{N_1}}+\delta \overset{③}{\leqslant} \frac{2L_0\|x^0-x_*\|_2}{\sqrt{N_1}}. \quad (2.40)$$

Неравенство ① имеет место ввиду μ-сильной выпуклости функции $f(x)$ (см. неравенство (1.14)), неравенство ② — ввиду оценки (2.35), а неравенство ③ — ввиду оценки $N_1 \leqslant L_0^2R_0^2/\delta^2$. Выберите N_1 из условия $R_1 = R_0/2$. Из неравенства (2.40) получите, что

$$N_1 \simeq \frac{64L_0^2}{(\mu^2R_1^2)}.$$

♦ В этом месте появляется проблема с практической реализацией схемы рестартов. Дело в том, что в такой реализации предписано сделать число итераций, зависящее явно от параметра μ, который, как правило, либо просто неизвестен, либо грубо оценён, не говоря уже о возможности локальной настройки метода на значение этого параметра, отвечающего текущему положению метода. В отличие от адаптивной настройки на гладкость задачи (см. § 5), на данный момент не известны общие способы настройки на параметр сильной выпуклости, лучшие, чем рестарты по этому неизвестному параметру [471, 495]: решаем задачу с $\mu = \mu_0$; если метод не сходится, то полагаем $\mu := \mu/2$, снова решаем и т. д., пока не получим сходимость. При таком подходе есть некоторые тонкости с детектированием сходимости. Несложно показать, что число дополнительных вычислений при этом увеличится не более чем в 8 раз [24]. Впрочем, в некоторых случаях можно более изящно решать отмеченную проблему [269, 280, 385, 496, 523, 532, 572]. Мы вернёмся к этому вопросу в конце § 5. ♦

Далее, после N_1 итераций, рестартуйте исходный метод и положите $x^0 := \bar{x}^{N_1}$. Определите $N_2 \leqslant L_0^2 R_1^2/\delta^2$ из условия $R_2 = \|\bar{x}^{N_2} - x_*\|_2 = R_1/2$. Получите, что

$$N_2 \simeq \frac{64 L_0^2}{\mu^2 R_2^2}$$

и т. д. После k таких рестартов общее число итераций будет составлять

$$N = N_1 + \ldots + N_k \simeq \frac{256 L_0^2}{\mu^2 R_0^2}(1 + 4^1 + \ldots + 4^{k-1}) \approx \frac{4^{k+4} L_0^2}{\mu^2 R_0^2}. \tag{2.41}$$

Обозначьте через $\tilde{x}^N$ то, что получается после N описанных итераций рестартованным методом. Обозначьте $\varepsilon = f(\tilde{x}^N) - f(x_*)$. Из неравенства (2.40) получите, что

$$\varepsilon = \frac{\mu R_k^2}{2} = \frac{2 L_0 R_{k-1}}{\sqrt{N_k}}. \tag{2.42}$$

Покажите, что из неравенства $N_k \leqslant L_0^2 R_{k-1}^2/\delta^2$ с учётом соотношения (2.42) следует, что $\varepsilon \geqslant 2\delta$. Покажите, что из оценки (2.41) следует соотношение

$$\frac{\mu R_k^2}{2} = \frac{128 L_0^2}{\mu N}. \tag{2.43}$$

Объединяя соотношения (2.42), (2.43) и неравенство $\varepsilon \geqslant 2\delta$, получите оценку (2.39).

Упражнение 2.4. Покажите, что прокс-функции, собранные в табл. 1, действительно 1-сильно выпуклы в соответствующих нормах.

Указание. См. [164, п. 5.6].

Упражнение 2.5. Предложите норму и прокс-функцию для задачи оптимизации на прямом произведении симплексов.

Указание. См. [29].

Упражнение 2.6 (адаптивные субградиентные методы [76, 158, 258, 481, 583]). Покажите, что для задачи негладкой выпуклой оптимизации (2.2) метод

$$x^{k+1} = x^k - h_k \nabla f(x^k)$$

при

$$h_k \equiv \frac{\varepsilon}{L_0^2}, \quad h_k \equiv \frac{R}{L_0\sqrt{N}}, \quad h_k \equiv \frac{\varepsilon}{\|\nabla f(x^k)\|_2^2}, \quad h_k \equiv \frac{R}{\|\nabla f(x^k)\|_2\sqrt{N}},$$

где $R = \|x_* - x^0\|_2$, будет сходиться согласно оценке (2.35) с

$$\bar{x}^N = \widehat{x}^N \in \operatorname{Arg}\min_{k=0,\ldots,N-1} f(x^k).$$

Покажите, что в общем случае решения задачи оптимизации на множестве простой структуры и использования неевклидовой нормы результат останется верным с заменой метода (2.2) *на метод зеркального спуска* (2.30) [158, 164, 186] и[5]

$$R = \sqrt{2V(x_*, x^0)}, \quad \|\nabla f(x^k)\|_2 \to \|\nabla f(x^k)\|_*.$$

Используя работы [158, 366, 570], предложите обобщение алгоритма (2.30) на задачи негладкой выпуклой оптимизации с негладкими выпуклыми функциональными ограничениями.

Указание (см. [4]). Для простоты рассмотрим здесь задачу безусловной оптимизации и ограничимся сначала случаем

$$h_k \equiv h = \frac{R}{L_0\sqrt{N}}.$$

Из структуры метода следует, что

$$\begin{aligned}\|x - x^{k+1}\|_2^2 &= \|x - x^k + h\nabla f(x^k)\|_2^2 = \\ &= \|x - x^k\|_2^2 + 2h\langle \nabla f(x^k), x - x^k\rangle + h^2\|\nabla f(x^k)\|_2^2 \leqslant \\ &\leqslant \|x - x^k\|_2^2 + 2h\langle \nabla f(x^k), x - x^k\rangle + h^2 L_0^2.\end{aligned}$$

[5] При аккуратном анализе, необходимом, например, для получения оптимальных оценок в приложении рассматриваемых методов к задаче о многоруких бандитах [30], [190, п. 5.2, 7.1], вместо $\|\nabla f(x^k)\|_*$ можно писать более точное выражение, см. также [583] и [364].

Отсюда (при $x = x_*$) получаем, что

$$f(\widehat{x}^N) - f(x_*) = \min_{k=0,\dots,N-1} f(x^k) - f(x_*) \leqslant$$
$$\leqslant \frac{1}{N}\sum_{k=0}^{N-1} f(x^k) - f(x_*) \leqslant \frac{1}{N}\sum_{k=0}^{N-1} \langle \nabla f(x^k), x^k - x_*\rangle \leqslant$$
$$\leqslant \frac{1}{2hN}\sum_{k=0}^{N-1} \left\{\|x_* - x^k\|_2^2 - \|x_* - x^{k+1}\|_2^2\right\} + \frac{hL_0^2}{2} =$$
$$= \frac{1}{2hN}\left(\|x_* - x^0\|_2^2 - \|x_* - x^N\|_2^2\right) + \frac{hL_0^2}{2}.$$

Полагая $h = R/(L_0\sqrt{N})$, получите неравенство

$$f(\widehat{x}^N) - f(x_*) \leqslant \frac{L_0 R}{\sqrt{N}}.$$

Заметим, что если функция $f(x)$ имеет L_1-липшицев градиент, то вместо оценки $\|\nabla f(x^k)\|_2^2 \leqslant L_0^2$, которая для негладких задач в общем случае не может быть улучшена по мере приближения к решению, можно использовать оценку (1.8)

$$\|\nabla f(x^k)\|_2^2 \leqslant 2L \cdot (f(x^k) - f(x_*)),$$

которая уже отражает уменьшение нормы градиента по мере приближения к решению. В результате при выборе шага $h = 1/(2L_1)$ получается следующая оценка (следует сравнить с оценкой (2.23)):

$$f(\widehat{x}^N) - f(x_*) \leqslant \frac{2L_1 R^2}{N}.$$

Для негладких задач выпуклой оптимизации на множествах простой структуры рассуждения будут аналогичными тем, что были приведены в § 2 при выводе формулы (2.11). Только для оценки $\mathrm{Prog}^h(x^k)$ придётся ограничиться неравенством Гёльдера (ввиду отсутствия липшицевости градиента)

$$\mathrm{Prog}^h(x^k) = \max_{x\in Q}\left\{\langle \nabla f(x^k), x^k - x\rangle - \frac{1}{2h}\|x^k - x\|_2^2\right\} \leqslant$$
$$\leqslant \max_{z}\left\{\langle \nabla f(x^k), z\rangle - \frac{1}{2h}\|z\|_2^2\right\} = \frac{h}{2}\|\nabla f(x^k)\|_2^2.$$

Простая структура приведённых выше рассуждений позволяет аналогичным образом провести рассуждения и в случае адаптивного выбора шага h_k. Ограничимся случаем

$$h_k \equiv \frac{R}{\|\nabla f(x^k)\|_2\sqrt{N}}.$$

В этом случае

$$\frac{R}{\sqrt{N}}\sum_{k=0}^{N-1}\left\langle\frac{\nabla f(x^k)}{\|\nabla f(x^k)\|_2},x^k-x_*\right\rangle\leqslant\frac{1}{2}\sum_{k=0}^{N-1}\left\{\|x_*-x^k\|_2^2-\|x_*-x^{k+1}\|_2^2\right\}+\frac{R^2}{2}\leqslant R^2.$$

Отсюда следует (см. упражнение 2.7), что

$$f(\widehat{x}^N)-f(x_*)\leqslant\frac{L_0}{N}\min_{k=0,\ldots,N-1}\left\langle\frac{\nabla f(x^k)}{\|\nabla f(x^k)\|_2},x^k-x_*\right\rangle\leqslant\frac{L_0R}{\sqrt{N}}.$$

Отметим, что адаптивность (суб-)градиентных методов можно получить за счёт других соображений, см. § 5. При этом подход, предложенный в § 5, является более универсальным, в частности потому, что позволяет работать с общей концепцией модели функции, рассмотренной в § 3. Отметим, что непонятно, как обобщать приведённые в данном упражнении адаптивные методы даже на частный случай общей модельной концепции: на задачи композитной оптимизации [18], см. пример 3.1.

Упражнение 2.7 (квазивыпуклые функции [76, п. 3.2.2–3.2.4], [80, п. 1.5], [337, 570]). Функция $f(x)$ называется квазивыпуклой на выпуклом множестве Q, если для любого $\alpha\in[0,1]$ и любых $x,y\in Q$ выполняется неравенство

$$f(\alpha x+(1-\alpha)y)\leqslant\max\{f(x),f(y)\}.$$

Покажите, что если функция $f(x)$ квазивыпуклая, то множества Лебега этой функции (множества вида $\{x\in Q\colon f(x)<C\}$) будут выпуклыми. Любая выпуклая функция будет квазивыпуклой, обратное в общем случае неверно.

Покажите (можно ограничиться случаем евклидовой нормы), что если $\nabla f(x)\neq 0$, то $f(x)-f(x_*)\leqslant\omega(\mathrm{v}(x))$, где модуль непрерывности определяется соотношением

$$\omega(t)=\max\left\{f(x)-f(x_*)\colon\|x-x_*\|\leqslant t\right\},\quad \mathrm{v}(x)=\frac{\langle\nabla f(x),x-x_*\rangle}{\|\nabla f(x)\|_*}.$$

(Здесь не приводится аккуратное определение $\nabla f(x)$. Чтобы лучше понять, какие на этом пути возникают сложности, можно вспомнить пример невыпуклой функции из § 1, с помощью которого получалась нижняя оценка сложности класса задач невыпуклой гладкой оптимизации. Описанная там функция была плохой с точки зрения локального оракула, выдающего локальную информацию о функции в запрошенной точке. В то же время данная функция является квазивыпуклой, а значит (согласно упражнению 2.7), вполне эффективно может

быть прооптимизирована. На первый взгляд кажется, что получается противоречие. Однако противоречия тут нет, поскольку эта функция во всех кубиках, кроме одного, тождественно равна нулю и, следовательно, $\nabla f(x) \equiv 0$ во всех таких кубиках. Таким образом, если понимать под $\nabla f(x)$, например, некоторый «перпендикуляр» к множеству Лебега, то это сразу приводит к нарушению условия локальности.)

Используя упражнение 2.6, предложите численный метод решения задач оптимизации с квазивыпуклым функционалом.

Указание. Ограничимся установлением неравенства

$$f(x) - f(x_*) \leqslant \omega(\mathrm{v}(x))$$

в евклидовом случае $\|\ \| = \|\ \|_2$ (в неевклидовом случае см. [158, 490]).

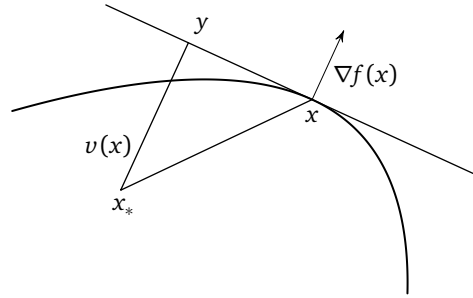

Рис. 8

На рис. 8 кривой линией изображена линия уровня функции $f(x)$. Обозначим через y проекцию точки x_* на касательную к этой линии уровня в точке x. По предположению квазивыпуклости функции $f(x)$ (множества Лебега выпуклые) точки x_* и y лежат по разные стороны от этой линии, поэтому

$$f(x) - f(x_*) \leqslant f(y) - f(x_*),$$

но последнюю величину ввиду определения функции $\omega(t)$ и того простого наблюдения, что в евклидовом случае расстояние между точками x_* и y равно $\mathrm{v}(x)$, можно оценить сверху величиной $\omega(\mathrm{v}(x))$.

В связи с полученным результатом интересно заметить, что $\mathrm{v}(x) \leqslant \|x - x_*\|$, поэтому неравенство $f(x) - f(x_*) \leqslant \omega(\mathrm{v}(x))$ является более тонким, чем просто условие липшицевой непрерывности функции $f(x)$.

Численный метод можно получить, если специальным образом выбирать шаг в субградиентном методе из упражнении 2.6:

$$h_k = \frac{R}{\|\nabla f(x^k)\|_* \sqrt{N}}.$$

§ 3

Общая схема получения оценок скорости сходимости. Структурная оптимизация

Как и в § 2, рассмотрим задачу выпуклой оптимизации (2.1):

$$f(x) \to \min_{x \in Q}.$$

Сначала обобщим условие (2.26) (см. также (2.3)). Будем говорить, что имеется (δ, L)-*модель функции* $f(x)$ в точке x (относительно нормы $\| \ \|$), и обозначать эту модель $(f_\delta(x); \psi_\delta(y, x))$, если для любого $y \in Q$ справедливо неравенство [97]

$$0 \leqslant f(y) - (f_\delta(x) + \psi_\delta(y, x)) \leqslant \frac{L}{2}\|y - x\|^2 + \delta, \tag{3.1}$$

где $\psi_\delta(y, x)$ — выпуклая функция по y, $\psi_\delta(x, x) = 0$, $\delta > 0$.

♦ Из неравенства (3.1) при $y = x$ следует, что $0 \leqslant f(x) - f_\delta(x) \leqslant \delta$, поэтому под (δ, L)-моделью функции $f(x)$ в точке x можно понимать только такую выпуклую по $y \in Q$ функцию $\psi_\delta(y, x)$, что для всех $y \in Q$ выполнено неравенство

$$f(x) + \psi_\delta(y, x) - \delta \leqslant f(y) \leqslant f(x) + \psi_\delta(y, x) + \frac{L}{2}\|y - x\|^2 + \delta. \ ♦$$

Частным случаем, отвечающим условиям

$$f_\delta(x) = f(x), \quad \psi_\delta(y, x) = \langle \nabla f(x), y - x \rangle, \tag{3.2}$$

такого определения является условие (2.26). Если не налагать условия $f_\delta(x) = f(x)$ в (3.2), то концепция (3.1), (3.2) совпадает с концепцией (δ, L)-оракула из работы [235], см. также [86, гл. 4], [150, 221, 554]. Близкие концепции модели функции имеются в работах [442, 497], см. также [145, 253]. Дальнейшее развитие данной концепции отражено в работах [42, 568, 569, 583].

♦ Из дальнейшего будет ясно, что в правой части неравенства (3.1) можно заменить $\|y - x\|^2$ на $2V(y, x)$. При этом правое неравенство в формуле (3.1) интерпретируют уже не как условие гладкости функции $f(x)$ (липшицевости градиента), а как *условие относительной*

гладкости [155, 439]. Важное преимущество, которое приобретается в случае такой замены, — отсутствие условия (2.31) на дивергенцию Брэгмана $V(y,x)$ (правда, и некоторые сложности приобретаются, например, задача (3.21) в таком случае может стать более сложной). Это наблюдение позволяет другим способом, отличным от описанного ранее в пособии, бороться с возможной неограниченностью параметра L, определяемого условиями (2.3) или (2.26), в случае неограниченного множества Q. Тут можно вспомнить пример $f(x)=x^4$, $Q=\mathbb{R}$ из § 2. С другой стороны, здесь, так же как и ранее в условии (2.3) (см. также (2.26)), достаточно потребовать, чтобы условие (3.1) выполнялось только для всех

$$x, y \in \{x \in Q : V(x_*, x) \leqslant R^2\},$$

где $R^2 = V(x_*, x^0)$, см. вторую половину § 2 и соотношения (3.16) ниже. Если решение не единственно, то в определении R^2 выбирается такое решения x_*, которое доставляет минимум R^2. ♦

Заметим, что неравенство (3.1) включает в себя намного больше свободы (см. [235]) по сравнению с (2.3). В частности, оно включает возможность неточного вычисления (суб-)градиента и значения функции, а не только игру на гладкости (см. соотношения (2.4), (2.5)). Мы вернёмся к более подробному обсуждению вопросов, связанных с концепцией (3.1), ниже, см. примеры 3.1, 3.2 и упражнения 3.2, 3.3, 4.3.

Подобно методу (2.29), рассмотрим следующий метод (пояснение записи (3.3) приведено ниже, в формуле (3.4)):

$$x^{k+1} = \arg_{\tilde{\delta}} \min_{x \in Q} \underbrace{\left\{ \psi_{\delta}(x, x^k) + \frac{1}{h} V(x, x^k) \right\}}_{\Psi(x, x^k)}, \tag{3.3}$$

где $V(x, x^k)$ — дивергенция Брэгмана, определённая в конце предыдущего параграфа. Если задача (3.3) точно решена, то существует такой

$$\nabla_{x^{k+1}} \Psi(x^{k+1}, x^k) \in \partial_x \Psi(x, x^k)\big|_{x=x^{k+1}},$$

что для любого $x \in Q$ выполнено неравенство

$$\left\langle \nabla_{x^{k+1}} \Psi(x^{k+1}, x^k), x - x^{k+1} \right\rangle \geqslant 0.$$

Однако мы будем допускать, что задача (3.3) решается лишь в следующем смысле:

$$\left\langle \nabla_{x^{k+1}} \Psi(x^{k+1}, x^k), x_* - x^{k+1} \right\rangle \geqslant -\tilde{\delta},$$

т. е. (следует сравнить с [150] и [164, п. 5.5.1.2])

$$\left\langle \nabla_{x^{k+1}} \Psi(x^{k+1}, x^k), x^{k+1} - x_* \right\rangle \leqslant \tilde{\delta}, \tag{3.4}$$

где $\tilde{\delta} > 0$. Добиться выполнения неравенства (3.4) можно по-разному, в зависимости от сложности задачи (3.3) (см. упражнение 3.1). Для возможности перенесения описанного в этом параграфе подхода на следующий параграф, другими словами, для обоснования прямодвойственности метода (3.3), необходимо отказаться от того, что $x = x_*$ в формуле (3.4). В этом случае нужно предполагать, что существует такой

$$\nabla_{x^{k+1}}\Psi(x^{k+1}, x^k) \in \partial_x \Psi(x, x^k)\big|_{x=x^{k+1}},$$

что

$$\max_{x\in Q}\left\langle \nabla_{x^{k+1}}\Psi(x^{k+1}, x^k), x^{k+1} - x\right\rangle \leqslant \tilde{\delta}. \tag{3.5}$$

Введём $\mathrm{Prog}^h_{\psi,V}(x^k)$ (см. также соотношения (2.9), (2.11)):

$$\mathrm{Prog}^h_{\psi,V}(x^k) = -\Big(\psi_\delta(x^{k+1}, x^k) + \frac{1}{h}V(x^{k+1}, x^k)\Big). \tag{3.6}$$

Из выпуклости $\psi_\delta(x, x^k)$ по x, определения x^{k+1} (формула (3.3)) и тождества ① в формуле (2.30), подобно выводу неравенства (2.10), с учётом оценки (3.5) (или (3.4), в этом случае можно сразу положить в последующих выкладках $x = x_*$) получим

$$\begin{aligned}
-\tilde{\delta} &\leqslant \left\langle \nabla_{x^{k+1}}\Psi(x^{k+1}, x^k), x - x^{k+1}\right\rangle = \\
&= \left\langle \nabla_{x^{k+1}}\psi_\delta(x^{k+1}, x^k) + \frac{1}{h}\nabla_{x^{k+1}}V(x^{k+1}, x^k), x - x^{k+1}\right\rangle = \\
&= \left\langle \nabla_{x^{k+1}}\psi_\delta(x^{k+1}, x^k), x - x^{k+1}\right\rangle + \frac{1}{h}V(x, x^k) - \\
&\quad - \frac{1}{h}V(x, x^{k+1}) - \frac{1}{h}V(x^{k+1}, x^k) \leqslant \psi_\delta(x, x^k) - \psi_\delta(x^{k+1}, x^k) + \\
&\quad + \frac{1}{h}V(x, x^k) - \frac{1}{h}V(x, x^{k+1}) - \frac{1}{h}V(x^{k+1}, x^k).
\end{aligned} \tag{3.7}$$

Отсюда следует, что

$$-\psi_\delta(x, x^k) \leqslant \mathrm{Prog}^h_{\psi,V}(x^k) + \tilde{\delta} + \frac{1}{h}V(x, x^k) - \frac{1}{h}V(x, x^{k+1}). \tag{3.8}$$

Согласно неравенству (3.1) при $y = x = x^k$ имеем

$$0 \leqslant f(x^k) - f_\delta(x^k) \leqslant \delta. \tag{3.9}$$

Отсюда по левому неравенству (3.1) при $y = x$, $x = x^k$ получаем

$$f(x^k) - f(x) - \delta \leqslant f_\delta(x^k) - f(x) \leqslant -\psi_\delta(x, x^k). \tag{3.10}$$

При $h \leqslant 1/L$ из соотношения (3.6) получаем

$$\begin{aligned}
\mathrm{Prog}^h_{\psi,V}(x^k) &= -\Big(\psi_\delta(x^{k+1}, x^k) + \frac{1}{h}V(x^{k+1}, x^k)\Big) \overset{①}{\leqslant} \\
&\overset{①}{\leqslant} f_\delta(x^k) - f(x^{k+1}) + \delta \overset{②}{\leqslant} f(x^k) - f(x^{k+1}) + \delta.
\end{aligned} \tag{3.11}$$

Неравенство ① следует из соотношения (2.31) и правого неравенства (3.1) при $x = x^k$, $y = x^{k+1}$, неравенство ② следует из оценки (3.9). Подставляя неравенство (3.11) в формулу (3.8), получим

$$-\psi_\delta(x, x^k) \leqslant f(x^k) - f(x^{k+1}) + \tilde{\delta} + \delta + \frac{1}{h}V(x, x^k) - \frac{1}{h}V(x, x^{k+1}). \quad (3.12)$$

Подставляя неравенство (3.10) в формулу (3.12), получим аналог неравенства (2.15):

$$f(x^k) - f(x) \leqslant f(x^k) - f(x^{k+1}) + \tilde{\delta} + 2\delta + \frac{1}{h}V(x, x^k) - \frac{1}{h}V(x, x^{k+1}),$$

т. е. при $h \leqslant 1/L$ имеет место основное неравенство

$$f(x^{k+1}) - f(x) \leqslant \frac{1}{h}V(x, x^k) - \frac{1}{h}V(x, x^{k+1}) + \tilde{\delta} + 2\delta. \quad (3.13)$$

Мы остановимся на этой формуле, так как все дальнейшие рассуждения в точности совпадают с аналогичными рассуждениями из предыдущего параграфа. Общий вывод, который можно сделать из неравенства (3.13), сформулируем следующим образом.

Теорема 3.1. *Пусть нужно решить задачу* (2.1). *Для метода* (3.3), (2.21), *т. е. для*

$$x^{k+1} = \arg_{\tilde{\delta}} \min_{x \in Q} \left\{ \psi_\delta(x, x^k) + LV(x, x^k) \right\}, \quad (3.14)$$

в условиях (3.1), (3.4) *имеют место оценки, аналогичные оценкам*[1] (2.22), (2.19):

$$f(\bar{x}^N) - f(x_*) \leqslant \frac{LR^2}{N} + \tilde{\delta} + 2\delta, \quad (3.15)$$

где

$$\bar{x}^N = \frac{1}{N}\sum_{k=1}^{N} x^k, \quad V(x_*, x^k) \leqslant V(x_*, x^0), \quad (3.16)$$

$R^2 = V(x_*, x^0)$. *Если решение x_* не единственно, то оценки* (3.15), (3.16) *будут верны для того решения x_*, которое доставляет минимум R^2.*

Рассмотрим пару примеров задач *структурной оптимизации* [75], демонстрирующих полезность рассмотрения более общих ситуаций, чем (3.2).

Пример 3.1 (композитная оптимизация). Рассмотрим следующую задачу *композитной оптимизации* (composite optimization) [162, 471]:

$$f(x) = F(x) + g(x) \to \min_{x \in Q} \quad (3.17)$$

[1] С оговоркой, аналогичной (2.18), в случае оценки (3.16).

с выпуклой функцией $F(x)$, удовлетворяющей условию (2.3), и, вообще говоря, негладкой выпуклой функцией $g(x)$ простой структуры. Последнее означает, что множества Лебега

$$\Lambda_y = \{x \in Q : g(x) < y\} \tag{3.18}$$

имеют простую структуру. К такой задаче, например, можно отнести задачу LASSO:

$$\frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1 \to \min_{x\in\mathbb{R}^n}.$$

Естественным обобщением метода (2.29) для задачи (3.17) будет следующий метод:

$$x^{k+1} = \arg\min_{x\in Q}\left\{\langle \nabla F(x^k), x-x^k\rangle + g(x) + LV(x,x^k)\right\}. \tag{3.19}$$

Метод (3.19) в точности соответствует методу (3.14) с

$$\psi_\delta(y,x) = \langle \nabla F(x), y-x\rangle + g(y) - g(x). \tag{3.20}$$

Таким образом, все приведённые выше результаты удаётся полностью перенести на задачи композитной оптимизации (3.17). В частности, имеет место оценка скорости сходимости (2.22). Стоит особо подчеркнуть, что в полученную оценку скорости сходимости никак не вошла информация о композите $g(x)$. Это может показаться странным, однако всё становится на места, если заметить, что по принципу множителей Лагранжа [66, п. 2.1], использованному в «обратном направлении», при весьма общих условиях существует такое y, что задача (3.17) эквивалентна задаче

$$F(x) \to \min_{x\in\Lambda_y}$$

с множеством Λ_y (см. формулу (3.18)) простой структуры[2].

Взяв в качестве композитного члена индикаторные функции выпуклых множеств простой структуры, можно получить результаты § 2 из композитного подхода с $Q = \mathbb{R}^n$.

Взяв в качестве композитного члена линейные функции, несложно понять, что скорость сходимости метода (3.19) в негладком случае (см. соотношения (2.3)–(2.5) с $\nu = 0$) зависит от константы L_0, а не от константы Липшица оптимизируемого функционала [162, 479]. Это можно понять непосредственно из самой оценки (2.22), т. е. без композитной оптимизации. Однако с композитной оптимизацией это свойство становится более ясным. ■

[2] Есть и другой способ, объясняющий факт отсутствия в оценке скорости сходимости информации о $g(x)$ [20, замечание 6].

В связи с примером 3.1 можно заметить, что если для обычной (некомпозитной) задачи, вообще говоря, негладкой выпуклой оптимизации (2.1) $f(x) \to \min_{x \in Q}$ взять в описанном выше подходе (для простоты считаем, что $\delta = 0$)

$$\psi_\delta(y, x) = f(y) - f(x)$$

и выбрать произвольное L в условии (3.1), то полученный по формуле (3.3) метод

$$x^{k+1} = \arg\min_{x \in Q}\{f(x) + LV(x, x^k)\} \tag{3.21}$$

становится известным *прокс-методом* решения задачи (2.1) [86, § 1, гл. 6], [26, 156, 180, 206, 241, 352, 375, 476], см. также замечание 3.2 в случае евклидовой прокс-структуры. Метод (3.21) будет сходиться согласно оценке (3.15) из теоремы 3.1, т. е. быстрее, чем следует ожидать исходя из нижней оценки, см. упражнение 2.1. Проблема, однако, в том, что в оценку (3.15) входит $\tilde{\delta}$ — «точность» решения вспомогательной задачи. Согласно упражнениям 2.3, 3.1 сложность решения вспомогательной задачи, которую можно понимать уже как задачу композитной оптимизации с L-сильно выпуклым композитом[3] $LV(x, x^k)$, будет не меньше чем $\widetilde{\mathrm{O}}(L_0^2/(L\tilde{\delta}))$, где L_0 определяется по формуле (2.4)[4]. Здесь под сложностью понимается число вычислений $\nabla f(x)$ и число решений уже стандартных вспомогательных подзадач вида (2.29). Комбинируя оценку (3.15) с оценкой $\mathrm{O}(L_0^2/(L\tilde{\delta}))$ и выбирая $\tilde{\delta} \sim \varepsilon$, где ε — желаемая точность (по функции) решения исходной задачи (2.1), получим оценку вида (2.34), что уже соответствует нижней оценке (2.36). Действительно, выбирая N в формуле (3.15) из условия $LR^2/N \sim \varepsilon$, получим для итоговой сложности соотношение

$$N\frac{L_0^2}{L\tilde{\delta}} \sim \frac{LR^2}{\varepsilon}\frac{L_0^2}{L\varepsilon} = \frac{L_0^2R^2}{\varepsilon^2}, \tag{3.22}$$

[3] То, что композит сильно выпуклый, нужно только для последующего обоснования градиентного слайдинга на базе прокс-метода. В теореме 3.1, описывающей, в частности, работу прокс-метода, достаточно только, чтобы $V(y, x)$ было дивергенцией Брэгмана, т. е. имело место представление

$$V(y, x) = d(y) - d(x) - \langle \nabla d(x), y - x\rangle.$$

.

[4] Строго говоря, в упражнении 2.3 рассматривается не композитная постановка, однако ввиду примера 3.1 несложно перенести результаты данного упражнения и на композитную постановку. Необходимые рассуждения, дословно повторяющие написанное в указании к упражнению 2.3, было решено здесь опустить, детали см., например, в [32, п. 2.3]. Также важно отметить, что задача (3.21) должна решаться с точностью $\tilde{\delta}$ в более сильном смысле (3.4), чем «по функции».

что соответствует оценке (2.34), если приравнять правую часть равенства (3.22) значению N из формулы (2.34) и выразить $\varepsilon(N)$.

♦ На первый взгляд кажется, что нет никакой выгоды от описанного в предыдущем абзаце подхода. Однако выгода получается в случае, когда приведённую конструкцию используют для задач с более сложной структурой, например для задач композитной оптимизации (3.17), но уже без предположения о простой структуре негладкой выпуклой функции $g(x)$, как в примере 3.1. В случае $V(x,y)=\frac{1}{2}\|x-y\|_2^2$ установлено, что исходную задачу (3.17) можно решить с точностью по функции ε за $\mathrm{O}(L_{0,g}^2R^2/\varepsilon^2)$ обращений к оракулу за субградиентом $g(x)$, где $L_{0,g}$ определяется согласно соотношению (2.4) с $\nu=0$ и $f\equiv g$, и $\mathrm{O}(L_{1,F}R^2/\varepsilon)$ обращений к оракулу за градиентом $F(x)$, где $L_{1,F}$ определяется согласно соотношению (2.4) с $\nu=1$ и $f\equiv F$. Удалось как бы «расщепить» задачу (3.17) на две задачи, отвечающие отдельным слагаемым, и организовать процедуру решения исходной задачи таким образом, чтобы сложность этой процедуры соответствовала суммарной сложности решения отдельных подзадач. В случае, когда вычисление градиента функции $F(x)$ занимает намного больше времени, чем вычисление субградиента функции $g(x)$, такое расщепление даёт очевидные преимущества [367, 404]. Приём, с помощью которого удаётся достичь описанного результата, называется *градиентным слайдингом* [2, 26, 42, 355, 374, 403, 404]. По-видимому, впервые он был предложен в работе [367]. В последние годы этот приём стал достаточно популярным в связи с большим числом приложений в задачах анализа изображений. За многочисленными обобщениями и приложениями данного приёма можно следить, например, по работам Дж. Лана [410]. Отметим, в частности, что схема слайдинга переносится и на функционалы, состоящие из суммы двух гладких композитов (см. [403, гл. 8] и упражнение 3.8), что также находит многочисленные приложения [2].

К сожалению, в общем случае техника слайдинга требует довольно тонких и весьма громоздких рассуждений для своего обоснования. Пожалуй, это единственная известная нам и достаточно широко использующаяся конструкция в современной выпуклой оптимизации, суть которой пока так и не удалось раскрыть (в случае, когда целевая функция представляется в виде суммы гладкого и негладкого слагаемого) с помощью элементарных соображений. Впрочем, популярное изложение более частного результата всё же имеется [571]. Для всех остальных основных конструкций в данном пособии предпринима-

ется попытка представить их достаточно простым (естественным) способом. ♦

Заметим также, что в работах [254, 433, 434] на базе *проксимального подхода* (описанного выше, см. формулу (3.21)) был предложен новый общий способ ускорения различных неускоренных методов, получивший название *катализт* (Catalyst). Немного подробнее об этом будет написано в замечании 3.3 и приложении.

Пример 3.2 (метод уровней). На пример 3.1 можно посмотреть и с немного другой точки зрения. Как уже отмечалось в самом начале § 2, типично имеется большой зазор между сложностью выполнения итерации $\widetilde{O}(n)$ (сложностью проектирования) и сложностью вычисления градиента[5] $O(n^2)$. Можно заметить, что простота множества Λ_y (композитной функции) на самом деле в рассуждениях примера 3.1 никак не использовалась. Она была нужна, чтобы не задумываться о сложности проектирования. Поэтому можно понимать пример 3.1 как способ (аддитивного) перенесения части сложности задачи в итерацию, благо для этого имеется хороший запас. Ведь всё равно, чтобы сделать шаг метода, нужно посчитать градиент, поэтому сложность «проектирования» вполне можно «утяжелять», например за счёт отмеченной идеи композитной оптимизации, до сложности расчёта градиента. Общая сложность итерации по порядку сохранится, но зато число итераций может существенно уменьшиться. Продолжая движение в намеченном направлении, приведём другой пример задачи «со структурой», которая также позволяет заносить часть сложности задачи в «проектирование», сохраняя общую конструкцию [71], [74, § 4, гл. 7], [76, п. 4.3], [402, п. 4], [462]:

$$f(x) = F(f_1(x), \ldots, f_m(x)) \to \min_{x \in Q}, \tag{3.23}$$

где все функции выпуклые, причём функция $F(y)$ ещё и неубывающая по каждому из своих аргументов. Также предполагаем, что все функции $f_j(x)$, $j = 1, \ldots, m$, удовлетворяют условию (2.3) с $L = L_j$, $j = 1, \ldots, m$, а функция $F(y)$ удовлетворяет условиям (2.3)–(2.5) с $\nu = 0$ ($L = L_0$) и $\| \ \| = \| \ \|_1$. Подобно соотношениям (3.2), (3.20) положим

$$\psi_\delta(y, x) = F\big(f_1(x) + \langle \nabla f_1(x), y - x\rangle, \ldots, f_m(x) + \langle \nabla f_m(x), y - x\rangle\big) - f(x). \tag{3.24}$$

[5] Такой большой зазор ($n \longleftrightarrow n^2$) имеет место не всегда. Однако в типичных ситуациях вычисление градиента занимает значительно больше времени, чем последующее проектирование.

Сделанные предположения позволяют утверждать, что условие (3.1) выполняется при $L = L_0 \sum_{j=1}^{m} L_j$. Получаемый при таком выборе функции $\psi_\delta(y, x)$ (см. формулу (3.24)) метод (3.3) называют *методом уровней* (level method).

Достаточно популярным частным случаем задачи (3.23) является задача, в которой $F(y) = \max_{j=1,\dots,m} y_j$ [76, п. 2.3]. К такой задаче с помощью *метода нагруженного функционала* [13, § 19, гл. 5], [86, § 3, гл. 9] сводятся и задачи *условной оптимизации* (задачи с функциональными ограничениями) вида

$$f_0(x) \to \min_{\substack{f_1(x)\leqslant 0,\dots,f_m(x)\leqslant 0,\\ x\in Q}}. \tag{3.25}$$

Действительно, задачу (3.25) можно переписать следующим образом: найти такой вектор $t = t_*$ и соответствующий вектор $x(t_*)$, доставляющий решение вспомогательной задачи минимизации в формуле (3.26), что $G(t) > 0$ при $t < t_*$ и $G(t_*) = 0$, где

$$G(t) = \min_{x\in Q} \max\{f_0(x) - t, f_1(x), \dots, f_m(x)\}. \tag{3.26}$$

Очевидно, что $G(t)$ — невозрастающая функция. Чуть посложнее показывается, что $G(t)$ — выпуклая функция.

Замечание 3.1. Это следует из двух общих фактов выпуклого анализа [182, гл. 3].

1. Пусть $\widetilde{F}(x, y)$ — выпуклая функция как функция переменной x, тогда функция

$$f(x) = \max_{y\in Q} \widetilde{F}(x, y)$$

также выпуклая. Хорошей иллюстрацией тут является представление $|x| = \max\{-x, x\}$, $x \in \mathbb{R}$.

2. Пусть $\overline{F}(x, y)$ — выпуклая функция как функция переменных (x, y), а $\overline{Q}$ — выпуклое множество, тогда функция

$$f(x) = \min_{y:\,(x,y)\in\overline{Q}} \overline{F}(x, y)$$

также выпуклая. Это следует из того, что пересечение надграфика выпуклой функции с выпуклым цилиндром с основанием Q также является выпуклым множеством и его проекция вдоль y также является выпуклым множеством. ■

Из общих результатов о поиске корня скалярного нелинейного уравнения [185, гл. 4], [477, Appendix A1] можно пытаться найти t_* с относительной точностью ε за $O(\ln \varepsilon^{-1})$ вычислений значения $G(t)$.

Каждое такое вычисление приводит к необходимости решения задачи вида (3.23). Поскольку задачу (3.23) можно решить в общем случае только приближённо, посчитать[6] $x(t)$ тоже можно только приближённо. Это обстоятельство приводит к необходимости более тонкого анализа. Детали см., например, в [76, п. 2.3]. Однако сохраняется общий вывод о возрастании сложности решения задачи (3.25) по сравнению с (3.23) в $O(\ln \varepsilon^{-1})$ раз при рассматриваемом подходе.

Заметим, что в случае, когда задача (3.25) негладкая, существуют и другие эффективные численные способы её решения, базирующиеся на *методе зеркального спуска* (см. упражнение 2.6) и замечании 4.3 и на методе зеркального спуска с переключениями, см., например, [158]. ■

Все последующие рассуждения могут проводиться в общности, выбранной в данном параграфе. Однако в методических целях далее мы намеренно не будем «гнаться за общностью» и будем стараться формулировать результаты таким образом, чтобы подчеркнуть в первую очередь обсуждаемую идею.

Упражнение 3.1. Пусть для задачи выпуклой оптимизации

$$f(x) \to \min_{x \in Q}$$

найдено ε-приближённое по функции решение $x_\varepsilon \in Q$, т. е.

$$f(x_\varepsilon) - f(x_*) \leqslant \varepsilon.$$

1. Пусть функция $f(x)$ удовлетворяет условию (2.27) при $\nu = 1$ с константой L_1 и $\nabla f(x_*) = 0$. Покажите, что тогда для всех $x \in Q$ имеет место следующая оценка:

$$\langle \nabla f(x_\varepsilon), x_\varepsilon - x \rangle \leqslant \|x_\varepsilon - x\| \sqrt{2L_1 \varepsilon}.$$

Пусть

$$R = \max_{x,y \in Q} \|y - x\| < \infty.$$

Покажите, что тогда имеет место следующая оценка:

$$\max_{x \in Q} \langle \nabla f(x_\varepsilon), x_\varepsilon - x \rangle \leqslant R \sqrt{2L_1 \varepsilon}.$$

2. Пусть функция $f(x)$ удовлетворяет на отрезке, соединяющем точки x_ε и x_*, условию (2.27) при $\nu = 1$ с константой L_1 и является

[6] Здесь $x(t)$ — решение вспомогательной задачи минимизации по $x \in Q$; см. формулу (3.26).

μ-сильно выпуклой в норме $\|\ \|$. Покажите, что тогда для всех $x \in Q$ имеет место следующая оценка:

$$\langle \nabla f(x_\varepsilon), x_\varepsilon - x\rangle \leqslant \left(L_1\|x_\varepsilon - x\| + \|\nabla f(x)\|_*\right)\sqrt{\frac{2\varepsilon}{\mu}}$$

и, как следствие,

$$\max_{x\in Q}\langle \nabla f(x_\varepsilon), x_\varepsilon - x\rangle \leqslant \left(L_1 R + \|\nabla f(x_*)\|_*\right)\sqrt{\frac{2\varepsilon}{\mu}}.$$

Упражнение 3.2. Пусть

$$f(x) = \min_{y\in Q} \bar{F}(y, x),$$

где $\widetilde{Q}$ — ограниченное выпуклое множество, а $\bar{F}(y, x)$ — такая достаточно гладкая выпуклая по совокупности переменных функция, что при $y, y' \in \widetilde{Q}$, $x, x' \in \mathbb{R}^n$ выполнено неравенство

$$\left\|\nabla\bar{F}(y', x') - \nabla\bar{F}(y, x)\right\|_2 \leqslant L\|(y', x') - (y, x)\|_2.$$

Пусть для произвольного x можно найти такой вектор $\tilde{y}_\delta(x) \in \widetilde{Q}$, что (следует сравнить с оценкой (3.4))

$$\max_{y\in Q}\left\langle \nabla_y\bar{F}(\tilde{y}_\delta(x), x), \tilde{y}_\delta(x) - y\right\rangle \leqslant \delta.$$

Покажите, что для любых $x, x' \in \mathbb{R}^n$ выполнены неравенства

$$\bar{F}(\tilde{y}_\delta(x), x) - f(x) \leqslant \delta, \quad \|\nabla f(x') - \nabla f(x)\|_2 \leqslant L\|x' - x\|_2$$

и

$$\left(\bar{F}(\tilde{y}_\delta(x), x) - 2\delta;\ \left\langle \nabla_y\bar{F}(\tilde{y}_\delta(x), x), y - x\right\rangle\right)$$

будет $(6\delta, 2L)$-моделью для функции $f(x)$ в точке x относительно 2-нормы.

Указание. См. [20]. Интересно сопоставить это упражнение с леммой 13 из книги [86, п. 5, § 1, гл. 5].

Упражнение 3.3 (прокс-метод с неточным решением задачи минимизации на итерации). Рассмотрим функцию (следует сравнить с задачей (3.21))

$$f_L(x) = \min_{y\in Q} \underbrace{\left\{f(y) + \frac{L}{2}\|y - x\|_2^2\right\}}_{\Psi(y,x)}.$$

Предположим, что $f(y)$ — выпуклая функция и

$$\max_{y\in Q}\left\{\Psi(y(x), x) - \Psi(y, x) + \frac{L}{2}\|y - y(x)\|_2^2\right\} \leqslant \delta.$$

Покажите, что тогда

$$\Big(f(y(x)) + \frac{L}{2}\|y(x) - x\|_2^2 - \delta;\, \langle L \cdot (x - y(x)), y - x\rangle\Big)$$

будет (δ, L)-моделью функции $f_L(x)$ в точке x относительно 2-нормы.

Указание. См. [235].

Замечание 3.2 (прокс-метод и сглаживание по Моро — Иосиде [425]). Введём функции

$$F_{L,x}(y) = f(y) + \frac{L}{2}\|y - x\|_2^2,$$
$$f_L(x) = \min_{y \in Q} F_{L,x}(y) = F_{L,x}(y_L(x)).$$

Для любого $L \geqslant 0$ имеет место неравенство [338, Proposition 12.29]

$$f_L(x) \leqslant f(x) \leqslant f_L(x) + \frac{M^2}{2L},$$

где M — константа Липшица функции $f(x)$ в 2-норме, причём выпуклая функция $f_L(x)$ будет иметь L-липшицев градиент. Кроме того, согласно [86, теорема 5, п. 2, § 1, гл. 6]

$$x_* \in \operatorname{Arg}\min_x f_L(x) \quad \Big(= \operatorname{Arg}\min_{x \in Q} f_L(x)\Big) \quad \Rightarrow$$
$$\Rightarrow \quad x_* \in \operatorname{Arg}\min_{x \in Q} f(x), \quad f_L(x_*) = f(x_*).$$

Таким образом, вместо исходной задачи можно рассматривать (сглаженную по Моро — Иосиде) задачу

$$f_L(x) \to \min_{x \in \mathbb{R}^n}(\min_{x \in Q}).$$

На эту задачу можно смотреть как на обычную задачу гладкой выпуклой оптимизации. Согласно упражнению 1.3 сложность решения этой задачи с точностью ε по функции (число вычислений градиента $\nabla f_L(x)$, т. е. число раз, которое необходимо решить вспомогательную задачу) быстрым градиентным методом можно оценить следующим образом:

$$O\Big(\sqrt{\frac{LR^2}{\varepsilon}}\Big).$$

Чем меньше выбирается параметр L, тем эта оценка будет лучше, но при этом тем сложнее на каждой итерации решать вспомогательную задачу. Заметим, что

$$\nabla f_L(x) = L \cdot (x - y_L(x)).$$

Поэтому обычный градиентный метод будет иметь вид

$$x^{k+1} = x^k - \frac{1}{L} L \cdot (x^k - y_L(x^k)) = y_L(x^k).$$

Однако согласно упражнению 3.3 внешнюю задачу можно решать и быстрым градиентным методом, который работает с концепцией (δ, L)-модели функции, например методом подобных треугольников из упражнения 3.7, см. также замечание 3.3. ■

Упражнение 3.4 (градиентное отображение). Подход, изложенный в этом параграфе, является далеко не единственным способом получения части описанных в нём результатов. Удобным инструментом также является использование *градиентного отображения*, см., например, [75], [76, п. 2.3]. С помощью градиентного отображения обобщается (путём замены градиента на градиентное отображение) основной набор базовых формул, из которых выводятся все последующие оценки[7], см., например, [76, п. 2.2.3, 2.3.2]. Попробуйте получить собранные в § 3 результаты с помощью градиентного отображения.

Упражнение 3.5 (модель для невыпуклой функции). Предложите обобщение концепции модели функции (3.1), пригодное для работы с невыпуклыми функциями.

Указание. См. [23, 175, 273, 442, 497, 568].

Упражнение 3.6. В работе [486] в связи с изучением процессов, происходящих в ходе избирательных компаний, и в связи с изучением быстрых способов кластеризации многомерных данных предлагается искать решение следующей задачи выпуклой оптимизации:

$$f_\mu(x = (z, p)) = g(\underbrace{(z, p)}_{x}) + \mu \sum_{k=1}^{n} z_k \ln z_k + \frac{\mu}{2}\|p\|_2^2 \to \min_{z \in S_n(1),\, p \in \mathbb{R}^m_+}.$$

Введём норму

$$\|x\|^2 = \|(z, p)\|^2 = \|z\|_1^2 + \|p\|_2^2.$$

[7] В связи с этим тезисом отметим, что сложность задачи оптимизации гладкой/негладкой выпуклой/сильно выпуклой функции для рассматриваемого класса численных методов (1.33) равносильна сложности задачи оптимизации функции, удовлетворяющей лишь определённому (явно выписываемому и конечному!) набору условий, связывающих значения функции и её (суб-)градиента в генерируемых методом (1.33) точках [582]. Это наблюдение позволяет получать точные минимаксные оценки скорости сходимости различных итерационных процедур вида (1.33) [227, 247–250, 386, 387, 579–582], см. также замечание 1.5.

Убедитесь, что $\|\ \|$ действительно норма. Предположим, что

$$\|\nabla g(x_2) - \nabla g(x_1)\|_* \leqslant L\|x_2 - x_1\|,$$

где $L \leqslant \mu$. Покажите, что если

$$\psi_\delta\big(y = (y^z, y^p), x = (x^z, x^p)\big) = \\ = \langle \nabla g(x), y - x\rangle + \mu \sum_{k=1}^n y_k^z \ln y_k^z + \frac{\mu}{2}\|y^p\|_2^2 - \\ - \Big(\mu \sum_{k=1}^n x_k^z \ln x_k^z + \frac{\mu}{2}\|x^p\|_2^2\Big) - \Big(L \sum_{k=1}^n y_k^z \ln\Big(\frac{y_k^z}{x_k^z}\Big) + \frac{L}{2}\|y^p - x^p\|_2^2\Big),$$

то при $\mu \geqslant L$ функция $\psi_\delta(y, x)$ выпукла по y, $\psi_\delta(x, x) = 0$ и для любых $y, x \in S_n(1) \otimes \mathbb{R}_+^m$ выполнены неравенства

$$f_\mu(x) + \psi_\delta(y, x) \leqslant f_\mu(y) \leqslant \\ \leqslant f_\mu(x) + \psi_\delta(y, x) + 2L \sum_{k=1}^n y_k^z \ln\Big(\frac{y_k^z}{x_k^z}\Big) + \frac{2L}{2}\|y^p - x^p\|_2^2.$$

Заметим, что выпуклость или простота функции $g(x)$ здесь не требуется! Используя концепцию модели функции в условиях относительной гладкости, предложите численный способ решения исходной задачи.

Указание. Идея такой модели была заимствована из работ [121, 127, 567].

Упражнение 3.7 (метод подобных треугольников [31, 97, 114, 477, 568, 569]). Для задачи (2.1) рассмотрите следующий вариант быстрого (ускоренного) градиентного спуска с одной проекцией, работающий с моделью функции (3.1):

$$y^0 = z^0, \quad A_0 = \alpha_0 = 0, \quad \alpha_{k+1} = \frac{1}{2L} + \sqrt{\frac{1}{4L^2} + \alpha_k^2}, \quad A_{k+1} = A_k + \alpha_{k+1},$$

$$x^{k+1} = \frac{\alpha_{k+1} z^k + A_k y^k}{A_{k+1}},$$

$$z^{k+1} = \arg_{\tilde\delta} \min_{x \in Q} \big\{\alpha_{k+1}\psi_\delta(x, x^{k+1}) + V(x, z^k)\big\},$$

$$y^{k+1} = \frac{\alpha_{k+1} z^{k+1} + A_k y^k}{A_{k+1}}.$$

Покажите, что

$$f(y^N) - f(x_*) = \mathrm{O}\Big(\frac{LR^2}{N^2} + \frac{L\tilde\delta}{N} + N\delta\Big).$$

Полезно сравнить эту формулу с (3.15).

Покажите, что оценка скорости сходимости описанного выше метода не ухудшится, если на каждой итерации делать в конце дополнительное присваивание: в качестве x^{k+1} выбирать ту точку среди $\{y^{k+1}, u^{k+1}, x^{k+1}\}$, которая доставляет наименьшее значение целевой (минимизируемой) функции. Для задач безусловной оптимизации с простейшей моделью функции (3.2) покажите, что получившийся в результате быстрый градиентный метод будет *релаксационным*, т. е. на генерируемой таким методом последовательности точек $\{x^k\}_k$ целевая функция будет монотонно убывать, см. также [163].

По аналогии с (2.19) и (3.16) покажите, что при $\delta = 0$ и $\tilde{\delta} = 0$ (похожую оценку, см., например, в [31], [477, Remark 6.1.1])

$$\max\left\{V(x_*, x^k), V(x_*, y^k), V(x_*, z^k)\right\} \leqslant V(x_*, x^0).$$

Замечание 3.3 (каталист и оптимальные тензорные методы [25, 26, 241, 352, 434, 450]). Упражнение 3.7 позволяет строить ускоренный метод с моделью функции

$$\psi_\delta(y, x) = f(y) - f(x).$$

Однако получающиеся в результате вспомогательные задачи за счёт роста $\alpha_k \sim k$ с ростом номера итерации будут всё хуже и хуже обусловленными. Более эффективным представляется способ, базирующийся на упражнении 3.3 и замечании 3.2, см. также [241][8]. В таком случае вспомогательные задачи будут намного проще: их обусловленность не меняется с ростом номера итерации. Платой за это является: 1) евклидова прокс-структура (впрочем, см. [241]); 2) необходимость достаточно точно решать вспомогательные задачи и 3) итоговый критерий качества работы метода $f_L(x^N) - f_L(x_*)$ вместо желаемого $f(x^N) - f(x_*)$. Заметим, что

$$f_L(x^N) - f_L(x_*) = f_L(x^N) - f(x_*) \leqslant f(x^N) - f(x_*).$$

Если функция $f(x)$ имеет L_f-липшицев градиент, то проблема 2 отсутствует, поскольку вспомогательные задачи решаются в нужном смысле за линейное время. Если дополнительно известно, что $f(x)$ ещё и μ_f-сильно выпуклая функция, то $f_L(x)$ также будет сильно выпуклой функцией с константой [425]

$$\tilde{\mu}_f = \mu_f \frac{L}{\mu_f + L} \simeq \mu_f \quad (\mu_f \ll L),$$

[8] Далее в этом замечании 3.3 используются обозначения, введённые в замечании 3.2.

поэтому исчезает и проблема 3. В действительности проблему 3 можно решить и без предположения о сильной выпуклости за счёт выбора специального варианта ускоренного внешнего метода [434]. Более того, проблемы 2, 3 просто и не возникают, если «правильно» выбрать ускоренный внешний метод. Ниже в обозначениях замечания 3.2 при $Q = \mathbb{R}^n$ приводится вариант «правильно» ускоренного градиентного метода (следует сравнить с методом линейного каплинга (МЛК) из указания к упражнению 1.3 и замечания 1.6 в части выбора y^{k+1}).

Инициализация (метод Монтейро — Свайтера)

Задаём z^0, y^0, $A_0 = 0$.

Основной цикл

Подбираем L_{k+1} и y^{k+1} *так, что*

$$\left[\begin{aligned}
&a_{k+1} = \frac{1/L_{k+1} + \sqrt{1/L_{k+1}^2 + 4A_k/L_{k+1}}}{2},\\
&A_{k+1} = A_k + a_{k+1},\\
&x^{k+1} = \frac{A_k}{A_{k+1}} y^k + \frac{a_{k+1}}{A_{k+1}} z^k,\\
&\|\nabla F_{L_{k+1}, x^{k+1}}(y^{k+1})\|_2 \leqslant \frac{L_{k+1}}{2}\|y^{k+1} - x^{k+1}\|_2 \text{ // условие Монтейро — // Свайтера,}
\end{aligned}\right.$$

$z^{k+1} = z^k - a_{k+1}\nabla f(y^{k+1})$.

Для последовательности $\{x^k, y^k, z^k\}_{k=1}^N$, генерируемой методом Монтейро — Свайтера, справедливы следующие неравенства [450]:

$$\frac{1}{2}\|z^N - x_*\|_2^2 + A_N \cdot (f(y^N) - f(x_*)) + \frac{1}{4}\sum_{k=1}^N A_k L_k \|y^k - x^k\|_2^2 \leqslant \frac{1}{2}\|x^0 - x_*\|_2^2 = \frac{R^2}{2},$$

$$f(y^N) - f(x_*) \leqslant \frac{R^2}{2A_N}, \quad \|z^N - x_*\|_2 \leqslant R, \quad \sum_{k=1}^N A_k L_k \|y^k - x^k\|_2^2 \leqslant 2R^2.$$

Последние два неравенства являются следствием первого (см. также замечание 1.6). Можно также получить оценку

$$A_N \geqslant \frac{1}{4}\left(\sum_{k=1}^N \frac{1}{\sqrt{L_k}}\right)^2.$$

Условие Монтейро — Свайтера позволяет вместо точного решения $y_{L_{k+1}}(x^{k+1})$ вспомогательной задачи, для которого (напомним, см. за-

мечание 3.2, что $F_{L,x}(y) = f(y) + \frac{L}{2}\|y-x\|_2^2$)

$$\left\|\nabla F_{L_{k+1},x^{k+1}}(y_{L_{k+1}}(x^{k+1}))\right\|_2 = 0,$$

искать только такое решение y^{k+1}, что (см. также [26, 152])

$$\|y^{k+1} - y_{L_{k+1}}(x^{k+1})\|_2 \leqslant \frac{L_{k+1}}{3L_{k+1}+L_f}\|x^{k+1} - y_{L_{k+1}}(x^{k+1})\|_2$$

и, значит, [352],

$$\|\nabla F_{L_{k+1},x^{k+1}}(y^{k+1})\|_2 \leqslant \frac{L_{k+1}}{2}\|y^{k+1} - x^{k+1}\|_2,$$

откуда следует неравенство (неулучшаемое в общем случае с точностью до числового множителя)

$$\left\|\nabla f_{L_{k+1}}(x^{k+1}) - L_{k+1}\cdot(x^{k+1} - y^{k+1})\right\|_2 \leqslant \|\nabla f_{L_{k+1}}(x^{k+1})\|_2,$$

где

$$\nabla f_{L_{k+1}}(x^{k+1}) = L_{k+1}\cdot(x^{k+1} - y_{L_{k+1}}(x^{k+1})).$$

Последнее неравенство можно понимать так, что для задачи

$$f_{L_{k+1}}(x) \to \min_{x\in\mathbb{R}^n}$$

доступен зашумлённый градиент $L_{k+1}\cdot(x^{k+1} - y^{k+1})$ с относительными детерминированными помехами. В работе [86, п. 3, § 2, гл. 4] отмечается, что такие помехи (с точностью до числового множителя, меньшего единицы, в правой части неравенства) не меняют по порядку картину сходимости обычного градиентного спуска, см. также текст перед замечанием 1.1. Впрочем, в данном подходе используется ускоренный градиентный метод, поэтому приведённые рассуждения не следует воспринимать как доказательство.

♦ В описанном подходе параметр L можно выбирать по-разному на разных итерациях. Ограничимся сначала случаем, когда $L_k \equiv L$, и попробуем подобрать значение этого параметра L. Для этого предположим, что у нас есть некоторый метод, позволяющий решать задачи выпуклой оптимизации с целевым функционалом $f(x)$, обладающим L_f-липшицевым градиентом и являющимся μ_f-сильно выпуклым, со сложностью

$$\widetilde{\mathrm{O}}\left(\Xi\left(\frac{L_f}{\mu_f}\right)\right),$$

равной числу вычислений $\nabla f(x)$. Тогда «стоимость» каждой итерации метода Монтейро — Свайтера будет равна

$$\widetilde{\mathrm{O}}\left(\Xi\left(\frac{L_f+L}{\mu_f+L}\right)\right),$$

а число итераций составит

$$\widetilde{\mathrm{O}}\Big(\sqrt{\frac{L}{\tilde{\mu}_f}}\Big) = \widetilde{\mathrm{O}}\Big(\sqrt{\frac{L}{\mu_f}}\Big).$$

Таким образом, разумно подбирать значение параметра L из условия

$$\Xi\Big(\frac{L_f + L}{\mu_f + L}\Big)\sqrt{\frac{L}{\mu_f}} \to \min_{\mu_f \leqslant L \leqslant L_f}.$$

В частности, если

$$\Xi\Big(\frac{L_f}{\mu_f}\Big) = \mathrm{O}\Big(\frac{L_f}{\mu_f}\Big),$$

то следует выбрать $L \simeq L_f$. Тогда

$$\widetilde{\mathrm{O}}\Big(\Xi\Big(\frac{L_f + L}{\mu_f + L}\Big)\sqrt{\frac{L}{\mu_f}}\Big) = \widetilde{\mathrm{O}}\Big(\sqrt{\frac{L_f}{\mu_f}}\Big),$$

что соответствует ускоренному градиентному методу, см. § 1. Таким образом, с помощью указанной выше конструкции на базе неускоренного градиентного спуска можно построить ускоренный. Описанный общий способ ускорения неускоренных методов первого и нулевого порядка, т. е. использующих производные оптимизируемой функции и её значения, получил название *катализт* (*catalyst*) [433, 434]. Конструкция каталист переносится и на вариационные неравенства, см. замечание 5.1 и [199, п. 3], [503]. Имеется естественное обобщение данной конструкции на задачи стохастической оптимизации [399]. Содержательные примеры ускорения неускоренных рандомизированных методов каталистом будут приведены в упражнении 3.8 и приложении. ♦

Поскольку внешний метод в конструкции Монтейро — Свайтера первого порядка, может показаться, что описанная выше конструкция не в состоянии ускорять методы более высокого порядка. Так оно и есть, если считать параметр L фиксированным на итерациях. Однако чем меньше выбирается параметр L, тем оценка скорости сходимости внешнего метода

$$\mathrm{O}\Big(\sqrt{\frac{LR^2}{\varepsilon}}\Big)$$

будет лучше, но при этом тем сложнее на каждой итерации решать вспомогательную задачу, чтобы посчитать градиент $\nabla f_L(x)$ с нужной точностью. Идея, позволяющая применять подход Монтейро — Свайтера для ускорения методов высокого порядка, состоит в следующем.

1. Вместо задачи $f_L(x) \to \min_{x\in\mathbb{R}^n}$ с фиксированным L следует рассмотреть параметрическое семейство задач $f_{L_{k+1}}(x) \to \min_{x\in\mathbb{R}^n}$ со специальным образом убывающей (на внешних итерациях) последовательностью $\{L_{k+1}\}_k$. Все эти задачи оптимизации имеют одинаковое решение x_*, которое необходимо найти.
2. Для приближённого решения вспомогательной задачи

$$F_{L_{k+1},x^{k+1}}(y) \to \min_{y\in\mathbb{R}^n},$$

возникающей на каждой внешней итерации, используется всего одна итерация неускоренного (тензорного) метода p-го порядка:

$$y^{k+1} = T^{F_{L_{k+1},x^{k+1}}}_{p,pM_p}(x^{k+1}),$$

где

$$T^{F_{L,x}}_{p,pM_p}(x) = \arg\min_{y\in\mathbb{R}^n}\left\{\sum_{r=0}^{p}\frac{1}{r!}[\nabla^r_z F_{L,x}(z)]_{z=x}\underbrace{[y-x,\dots,y-x]}_{r} + \right.$$
$$\left. + \frac{pM_p}{(p+1)!}\|y-x\|_2^{p+1}\right\},$$

и предполагается, что (см. приложение)

$$\|\nabla^p f(y) - \nabla^p f(x)\|_2 \leqslant M_p\|y-x\|_2, \quad x,y\in\mathbb{R}^n,\ M_p\leqslant\infty.$$

Если для такого метода выбирать L_{k+1} так, чтобы выполнялось условие Монтейро — Свайтера и условие [25, 450]

$$\frac{2(p+1)}{p!}\frac{M_p}{L_{k+1}}\|y^{k+1}-x^{k+1}\|_2^{p-1} \geqslant \frac{1}{2},$$

то число внешних итераций (число вычислений оператора $T^{F_{L,x}}_{p,pM_p}(x)$, а следовательно, и $\{\nabla^r f(x)\}_{r=1}^p$) будет определяться оценкой

$$\mathrm{O}\left(\left(\frac{M_p R^{p+1}}{\varepsilon}\right)^{2/(3p+1)}\right),$$

неулучшаемой для данного класса задач на классе методов p-го порядка, см. приложение.

Из работы [473] следует, что для всех $x\in\mathbb{R}^n$ имеет место следующее неравенство:

$$\left\|\nabla F(T^F_{p,pM_p}(x))\right\|_2 \leqslant \frac{(p+1)M_p}{p!}\|T^F_{p,pM_p}(x)-x\|_2^p$$

Таким образом, L_{k+1} нужно подбирать из условий

$$\frac{1}{2} \leqslant \frac{2(p+1)}{p!}\frac{M_p}{L_{k+1}}\left\|T^{F_{L_{k+1},x^{k+1}}}_{p,pM_p}(x^{k+1})-x^{k+1}\right\|_2^{p-1} \leqslant 1.$$

Далее заметим, что при $x^{k+1} \neq x_*$ найдутся такое, вообще говоря, достаточно маленькое значение $\check{L}_{k+1} > 0$, что

$$\frac{2(p+1)}{p!}\frac{M_p}{\check{L}_{k+1}}\left\|T^{F_{\check{L}_{k+1},x^{k+1}}}_{p,pM_p}(x^{k+1}) - x^{k+1}\right\|_2^{p-1} \geqslant 1,$$

и такое достаточно большое значение $\bar{L}_{k+1} > 0$, что

$$\frac{2(p+1)}{p!}\frac{M_p}{\bar{L}_{k+1}}\left\|T^{F_{\bar{L}_{k+1},x^{k+1}}}_{p,pM_p}(x^{k+1}) - x^{k+1}\right\|_2^{p-1} \leqslant \frac{1}{2}.$$

Отсюда ввиду непрерывной зависимости $T^{F_{L_{k+1},x^{k+1}}}_{p,pM_p}(x^{k+1})$ от L_{k+1} следует, что подобрать L_{k+1} можно с помощью процедуры вида $L_{k+1} = 2L_k$, $L_{k+1} := L_{k+1}/\sqrt{2}$. Конечно, есть риск «проскочить» нужный диапазон. В таком случае можно предусмотреть процедуру «возврата», имеющую вид $L_{k+1} := \sqrt[4]{2}L_{k+1}$ и т. д. В типичных ситуациях можно ожидать, что число вызовов оператора $T^{F_{L,x}}_{p,pM_p}(x)$ на одной итерации внешнего метода будет составлять $O(1)$, см. также [189, 361]. При этом каждый вызов такого оператора порождает свою выпуклую задачу (то, что задача получается выпуклой, — нетривиальный факт, который был обнаружен совсем недавно [473, 475]). Сложность решения такой задачи (т. е. вычисление $T^{F_{L,x}}_{p,pM_p}(x)$) с нужной точностью сопоставима при $p = 2, 3$ по объёму вычислений со сложностью итерации метода Ньютона, т. е. оценивается как $\widetilde{O}(n^{2,37})$. Отметим, что при решении возникающей задачи при $p = 3$ используется концепция относительной гладкости и относительной сильной выпуклости (см. упражнение 3.10) [473, 475, 569], см. также начало § 3. В оценке $\widetilde{O}(n^{2,37})$ не учитывается время расчёта

$$[\nabla^r_z F_{L,x}(z)]_{z=x}\underbrace{[y-x,\ldots,y-x]}_{r}.$$

Приведённое выражение во многих интересных на практике случаях может быть эффективно посчитано с помощью автоматического дифференцирования [473]. Дополнительную информацию о тензорных методах можно найти, например, в приложении. ■

♦ Близкие к описанному выше подходу рассуждения имеются в работе [241]. Различие в том, что в ней в качестве внешнего ускоренного метода используется метод без вспомогательного одномерного поиска, что в конечном итоге приводит к чуть более плохим теоретическим оценкам оракульной сложности предложенных в работе [241] тензорных методов (методов высокого порядка). Данная ситуация

была исправлена в работе [474] (см. также [476]). Отметим, что ускоренный проксимальный метод из работы [233] может работать с неевклидовым проксом вида дивергенции Брэгмана (см. § 2) в предположении, что прокс-функция имеет ограниченную константу Липшица градиента на допустимом множестве. ♦

Упражнение 3.8 (гладкий/ускоренный слайдинг [403, п. 8.2], [355]). Рассмотрим следующую задачу:

$$f(x) + g(x) \to \min_x,$$

где $f(x)$ и $g(x)$ имеют L_f- и L_g-липшицевы градиенты в 2-норме, причём $L_f \leqslant L_g$, а функция $g(x)$ является μ-сильно выпуклой в 2-норме, причём $\mu \leqslant L_f$. Покажите, что для решения этой задачи с заданной точностью[9] достаточно $\widetilde{\mathrm{O}}\left(\sqrt{L_f/\mu}\right)$ вычислений $\nabla f(x)$ и $\widetilde{\mathrm{O}}\left(\sqrt{L_g/\mu}\right)$ вычислений $\nabla g(x)$. Попробуйте получить аналогичный результат в модельной общности.

Указание. Применим к рассмотренной задаче технику каталист[10]. Тогда вместо исходной задачи потребуется $\widetilde{\mathrm{O}}(\sqrt{L/\mu})$ раз решать задачу вида

$$f(x) + g(x) + \frac{L}{2}\|x - x^k\|_2^2 \to \min_x.$$

Последнюю задачу можно решать неускоренным композитным градиентным методом (см. пример 3.1), считая $g(x) + \frac{L}{2}\|x - x^k\|_2^2$ композитом. Число итераций такого метода будет совпадать с числом вычислений $\nabla f(x)$ и будет равно $\widetilde{\mathrm{O}}(L_f/(L+\mu))$, где мы считаем, что $\mu \leqslant L \leqslant L_f$. Но в условиях задачи не предполагалась проксимальная дружественность функции $g(x)$, поэтому возникающую на каждой итерации неускоренного композитного градиентного метода задачу вида

$$\langle \nabla f(\tilde{x}^l), x - \tilde{x}^l\rangle + \frac{L_f}{2}\|x - \tilde{x}^l\|_2^2 + g(x) + \frac{L}{2}\|x - x^k\|_2^2 \to \min_x,$$

[9] Неважно, с какой именно точностью. Эта точность будет входить в выражения под знаком логарифма в приведённых далее оценках, а для наглядности логарифмические сомножители в данном упражнении было решено опустить. Далее в указании к этому упражнению оговорки о точности решения возникающих подзадач также опускаются, поскольку всё это влияет только на логарифмические сомножители в итоговых оценках, которые опущены.

[10] Обойтись без этой техники не получается по тем же причинам (см. начало замечания 3.3), по которым из ускоренного метода, описанного в упражнении 3.7, не получается с помощью модельной общности получить ускоренный проксимальный метод с оптимальной оценкой скорости сходимости.

в свою очередь, необходимо будет решать. Для решения данной задачи можно использовать ускоренный композитный градиентный метод для задач сильно выпуклой оптимизации (см. [31, 471], а также упражнение 3.7 и конец § 5), считая

$$\frac{L_f}{2}\|x-\tilde{x}^l\|_2^2+\frac{L}{2}\|x-x^k\|_2^2$$

композитом. Число итераций такого метода будет составлять

$$\widetilde{\mathrm{O}}\left(\sqrt{\frac{L_g}{L_f+L+\mu}}\right).$$

Таким образом, общее число вычислений $\nabla g(x)$ будет составлять

$$\widetilde{\mathrm{O}}\left(\sqrt{\frac{L}{\mu}}\right)\cdot\left[\widetilde{\mathrm{O}}\left(\frac{L_f}{L+\mu}\right)\cdot\widetilde{\mathrm{O}}\left(\sqrt{\frac{L_g}{L_f+L+\mu}}\right)+1\right].$$

Выбирая параметр $L\in[\mu,L_f]$ так, чтобы последнее выражение было минимальным, получим (с учётом сделанных предположений $L_f\leqslant L_g$ и $\mu\leqslant L_f$), что $L\simeq L_f$. Следовательно, общее число вычислений $\nabla g(x)$ будет равно $\widetilde{\mathrm{O}}\left(\sqrt{L_g/\mu}\right)$.

♦ Заметим, что данное упражнение можно обобщить на случай, когда вместо $\nabla g(x)$ доступно только $g(x)$ [355]. Чтобы понять, как всё это сделать, рекомендуется ознакомиться с безградиентными методами в начале приложения и цитированной там литературой. Приведённые в упражнении результаты можно улучшить, если дополнительно известно, что функция $g(x)$ имеет представление в виде суммы функций [42, 355]. Тогда вместо ускоренного композитного градиентного метода для задач сильно выпуклой оптимизации можно использовать ускоренный композитный метод редукции дисперсии для задач сильно выпуклой оптимизации, см., например, [406] и замечание 1 в приложении. Описанный в упражнении 3.8 слайдинг переносится и на тензорные методы [374].

Отметим также, что в описанной в указании к упражнению 3.8 трехуровневой схеме слайдинга можно убрать промежуточный уровень, связанный с использованием неускоренного композитного градиентного спуска. Для этого следует использовать композитный вариант ускоренной проксимальной оболочки (с гладким композитом), см., например, [26, 374, 375]. ♦

Упражнение 3.9 (проксимальный метод Синхорна [270, 610]). В последнее время в различных приложениях часто встречается расстояние Монжа — Канторовича — Васерштейна [508] между двумя ве-

роятностными мерами. Вычисление такого расстояния для дискретных мер сводится к классической транспортной задаче линейного программирования (ЛП)

$$\sum_{i,j=1}^{n} c_{ij}x_{ij} \to \min_{\substack{\sum_{j=1}^{n} x_{ij}=l_i,\ i=1,\dots,n \\ \sum_{i=1}^{n} x_{ij}=w_j,\ j=1,\dots,n \\ x_{ij}\geqslant 0,\ i,j=1,\dots,n}},$$

где $\sum_{i=1}^{n} l_i = \sum_{j=1}^{n} w_i = 1$. Используя неускоренный прокс-метод (3.21) с $V(x,y) = \sum_{i,j=1}^{n} x_{ij}\ln(x_{ij}/y_{ij})$:

$$x^{k+1} = \arg \min_{\substack{\sum_{j=1}^{n} x_{ij}=l_i,\ i=1,\dots,n \\ \sum_{i=1}^{n} x_{ij}=w_j,\ j=1,\dots,n \\ x_{ij}\geqslant 0,\ i,j=1,\dots,n}} \left\{ \sum_{i,j=1}^{n} c_{ij}x_{ij} + L \sum_{i,j=1}^{n} x_{ij}\ln\left(\frac{x_{ij}}{x_{ij}^k}\right)\right\},$$

предложите способ решения транспортной задачи.

Предложите адаптивный способ подбора параметра L.

Указание. Следует учесть, что возникающую на каждой итерации прокс-метода задачу можно приближённо решать путём перехода к двойственной задаче и использования метода альтернативных направлений: двойственную функцию можно явно прооптимизировать по группе (двойственных) множителей Лагранжа, отвечающих ограничениям

$$\sum_{j=1}^{n} x_{ij} = l_i, \quad i = 1,\dots,n,$$

и при «замороженных» остальных множителях. То же самое можно проделать и по группе оставшихся множителей, отвечающих ограничениям

$$\sum_{i=1}^{n} x_{ij} = w_j, \quad j = 1,\dots,n.$$

Чередуя такие оптимизации, получим метод Синхорна (Синхорна — Брэгмана — Шелейховского), также называемый методом балансировки, который представляет собой метод альтернативных направлений[11]

[11] Метод альтернативных направлений в худшем случае сходится как градиентный спуск (в наилучшей норме, см. замечание 1.3 и работу [588]) с константой Липшица градиента, отвечающей наименьшей из констант Липшица градиента целевой функции по соответствующей группе переменных [161]. Этот результат недавно был перенесён и на (ускоренные) блочно-покомпонентные методы (см., например, приложение), в которых вместо

для двойственной задачи [161]. Про этот метод (в приложении к данной задаче) известно, что его трудоёмкость имеет вид [270, 285, 567]

$$n^2 \min\left\{\widetilde{O}\left(\frac{1}{L\tilde{\varepsilon}}\right),\ \widetilde{O}\left(\exp\left(\frac{C(n)}{L}\right)\right)\right\}.$$

Согласно теореме 3.1 внешний прокс-метод с $R^2 = O(\ln n^2)$ сойдётся за

$$O\left(\frac{LR^2}{\varepsilon}\right) = \widetilde{O}\left(\frac{L}{\varepsilon}\right)$$

итераций, где точность решения внутренней задачи $\tilde{\varepsilon}$ должна быть существенно выше точности решения исходной задачи: $\tilde{\varepsilon} \ll \varepsilon$. Таким образом, описанный выше проксимальный метод при оптимальном выборе L будет иметь трудоёмкость[12]

$$n^2\widetilde{O}\left(\frac{C(n)}{\varepsilon}\right),$$

где $C(n) \gg n$. В действительности на практике описанный метод работает заметно лучше [567, 568]. Отметим в этой связи, что наилучшие с точки зрения теоретических оценок способы решения исходной транспортной задачи [173, 358, 420, 517], имеющие сложность

$$n^2 \cdot \min\left\{\widetilde{O}\left(\frac{1}{\varepsilon}\right), \widetilde{O}(\sqrt{n})\right\},$$

и (рассмотренного в этом упражнении) её L-энтропийно регуляризованного варианта [119, 215], имеющие сложность $\widetilde{O}(n^2/L)$, пока далеки от практической эффективности. В частности, солверы, решающие транспортную задачу как задачу ЛП с помощью методов внутренней точки (см. текст после замечания 4 приложения), имеют практическую сложность $\widetilde{O}(n^3)$ [506, 508], такую же сложность имеют для этой задачи симплекс-метод [287, 577] и венгерский алгоритм для специальной подзадачи «о назначениях» [396], см. также указание к упражнению 1.4.

шага типа градиентного спуска в одном из блоков осуществляется явная оптимизация по соответствующим этому блоку переменным [239]. Также недавно были предложены (прямодвойственные) ускоренные варианты метода альтернативных направлений с m блоками [239, 323]. В теоретическом плане метод из работы [323] требует в m раз больше итераций, чем обычный ускоренный метод, но на практике работает заметно быстрее последнего.

[12] Заметим, что если рассматриваемую транспортную задачу решать с помощью энтропийной регуляризации, то согласно замечанию 4.1 нужно выбирать $L = \varepsilon/(2\ln n^2)$ (более тонкий анализ конкретного случая — энтропийной регуляризации транспортной задачи — имеется в работе [601]), что приведёт в итоге к оценке трудоёмкости метода Синхорна $\widetilde{O}(n^2/\varepsilon^2)$.

Способ подбора параметра L можно построить, например, на базе следующей идеи. На первой итерации прокс-метода стартуем с завышенной оценки L, решаем задачу, затем полагаем $L := L/2$ и перерешиваем задачу, и так до тех пор, пока не детектируем существенное увеличение (например, в 10 раз) сложности решения вспомогательной задачи энтропийно-линейного программирования по сравнению со стартовой сложностью. Найденное значение L можно использовать и на последующих итерациях проксимального метода. В качестве точки старта новой итерации такого метода можно выбирать решение вспомогательной задачи с предыдущей итерации.

Упражнение 3.10 (имплементируемость тензорных методов третьего порядка; Ю. Е. Нестеров, 2018 [473]). Рассмотрим шаг тензорного метода из замечания 3.3 при $p = 3$:

$$T_{3,3M_3}^{F_{L,x}}(x) = \arg\min_{y\in\mathbb{R}^n} \underbrace{\left\{\sum_{r=0}^{3}\frac{1}{r!}[\nabla_z^r F_{L,x}(z)]_{z=x}\underbrace{[y-x,\ldots,y-x]}_{r} + \frac{3M_3}{4!}\|y-x\|_2^4\right\}}_{f(x)} =$$

$$= \arg\min_{y\in\mathbb{R}^n}\Big\{F_{L,x}(x) + \langle\nabla F_{L,x}(x), y-x\rangle + \frac{1}{2}\langle\nabla F_{L,x}(x)(y-x), y-x\rangle + \\ + \frac{1}{6}\nabla F_{L,x}(x)[y-x, y-x, y-x] + \frac{M_3}{8}\|y-x\|_2^4\Big\}.$$

Используя $\nabla F_{L,x}(x)$, $\nabla F_{L,x}(x)$, предложите способ приближённого[13] поиска $T_{3,3M_3}^{F_{L,x}}(x)$.

Указание. Прежде всего заметим, что вычислять $\nabla F_{L,x}(x)$ нет необходимости, потому что в описываемом далее подходе используется только

$$\nabla F_{L,x}(x)[y-x, y-x] \approx \frac{\nabla F_{L,x}(x+\tau(y-x))(y-x) - \nabla F_{L,x}(x)(y-x)}{\tau}.$$

Далее введём условие относительной сильной выпуклости. Будем говорить, что $f(x)$ является μ-сильно выпуклой и L-гладкой функцией относительно дивергенции Брэгмана

$$V(y,x) = d(y) - d(x) - \langle\nabla d(x), y-x\rangle,$$

если для всех $x, y \in \mathbb{R}$ (можно обобщить данное определение и основные выводы и на случай неточной модели $f(x)$ [569]) выполняются неравенства

$$f(x) + \langle\nabla f(x), y-x\rangle + \mu V(y,x) \leqslant f(y) \leqslant f(x) + \langle\nabla f(x), y-x\rangle + LV(y,x).$$

[13] Условия на точность решения приводятся в работах [243, 374, 375, 474–476].

Достаточным условием для этого будет выполнение соотношений

$$\mu\nabla d(y) \prec \nabla f(y) \prec L\nabla d(y).$$

Описанный в формуле (3.21) градиентный спуск

$$x = \arg_{\tilde{\delta}} \min_{x\in\mathbb{R}^n} \left\{\langle \nabla f(x), x - x\rangle + LV(x,x)\right\}$$

будет сходиться для такой функции согласно формуле [473, 476, 569]

$$\min_{k=0,\dots,N} f(x) - f(x_*) \leqslant (L+\mu)V(x_*,x)\min\left\{\frac{1}{N}, \left(\frac{L-\mu}{L+\mu}\right)\right\} + \tilde{\delta}.$$

Для рассматриваемой в условиях упражнения задачи оптимизации (для простоты обозначений будем считать, что ищется $T^{F_{L,x}}_{3,3M_3}(0)$) в работе [473] была предложена следующая прокс-функция:

$$d(y) = \frac{1}{2}\langle \nabla f(0)y, y\rangle + \frac{M_3}{8}\|y\|_2^4.$$

Для такой прокс-функции было установлено [476], что

$$\mu = 1 - \frac{1}{\sqrt{2}}, \quad L = 1 + \frac{1}{\sqrt{2}}.$$

Отсюда следует, что решение задачи поиска $T^{F_{L,x}}_{3,3M_3}(0)$ по сложности эквивалентно (с точностью до логарифмического множителя от желаемой точности поиска $T^{F_{L,x}}_{3,3M_3}(0)$) задаче вида

$$\langle b, y\rangle + \frac{1}{2}\langle Ay, y\rangle + \lambda\|y\|_2^4 \to \min_{y\in\mathbb{R}^n},$$

которая, в свою очередь, эквивалентна задаче

$$\langle b, y\rangle + \frac{1}{2}\langle Ay, y\rangle \to \min_{\|y\|_2^4 \leqslant C(\lambda)},$$

или, что эквивалентно,

$$\langle b, y\rangle + \frac{1}{2}\langle Ay, y\rangle \to \min_{\|y\|_2^2 \leqslant \sqrt{C(\lambda)}},$$

или, что эквивалентно,

$$\langle b, y\rangle + \frac{1}{2}\langle Ay, y\rangle + \tilde{\lambda}(\lambda)\|y\|_2^2 \to \min_{y\in\mathbb{R}^n},$$

где поиск $\tilde{\lambda}(\lambda)$ может быть осуществлён за логарифмическое время от желаемой точности. Последняя задача имеет такую же сложность (не больше), как и итерация метода Ньютона (см. приложение). Таким образом, с точностью до квадрата логарифмического по желаемой точности множителя задача поиска $T^{F_{L,x}}_{3,3M_3}(x)$ по сложности эквивалентна выполнению итерации метода Ньютона. Более того, так же

как и в методе Ньютона, здесь используется информация о функции не выше второго порядка $\nabla F_{L,x}(x)$, $\nabla^2 F_{L,x}(x)$. Другими словами, при применении описанного здесь внутреннего метода второго порядка тензорный метод третьего порядка в такой реализации становится, по сути, методом второго порядка. Если все описанные здесь действия осуществить более аккуратно (с контролем точности), то всему этому можно придать вполне законченный вид [375, 474, 476]. Полученные в результате такого подхода методы были названы Ю. Е. Нестеровым *супербыстрыми тензорными методами второго порядка* [375, 474, 476]. Об этом также вкратце будет рассказано в приложении.

Упражнение 3.11 (статистический предобуславливатель для задач минимизации суммы выпуклых функций [343, 574]). Рассмотрим задачу

$$f(x) = \frac{1}{m}\sum_{l=1}^{m} f_l(x) \to \min_{x\in\mathbb{R}^n},$$

где $f_l(x)$ — μ-сильно выпуклые в 2-норме гладкие функции. Будем считать, что m — большое число. Предположим, что есть централизованная архитектура с $r \ll m$ узлами, см. текст после упражнения 4.7. Первый узел является центральным, т. е. связан со всеми остальными. Поместим в первый узел случайно отобранные $\widetilde{m}$ ($\widetilde{m} \ll m$) слагаемых из суммы. Остальные слагаемые распределим по остальным узлам. Обозначим соответствующую первому узлу нормированную подсумму через $\widetilde{F}(x)$. Рассмотрим градиентный спуск (3.21), выполняемый на центральном (первом) узле:

$$x = \arg_{\tilde{\delta}} \min_{x\in\mathbb{R}^n} \left\{\langle \nabla f(x^k), x - x^k\rangle + LV(x, x^k)\right\},$$

где

$$V(y,x) = d(y) - d(x) - \langle \nabla d(x), y - x\rangle, \quad d(x) = \widetilde{F}(x) + \frac{\gamma}{2}\|x\|_2^2, \quad \gamma > 0.$$

Каждая итерация такого градиентного спуска должна отвечать коммуникации центрального узла с остальными, чтобы в итоге получилось $\nabla f(x^k)$.

Покажите, что если $\|\nabla^2\widetilde{F}(x) - \nabla^2 f(x)\|_2 \leqslant \gamma$ (определение матричной 2-нормы, согласованной с векторной 2-нормой, см. в приложении), то

$$\frac{\mu}{\mu + 2\gamma}\nabla^2 d(x) \prec \nabla^2 f(x) \prec \nabla^2 d(x).$$

Из матричного неравенства Хёфдинга с вероятностью $\geqslant 1 - \sigma$ имеем, что $\gamma \sim \sqrt{\ln(n/\sigma)/\widetilde{m}}$ [586]. Используя это наблюдение и указа-

ние к упражнению 3.10, оцените скорость сходимости градиентного спуска (3.21).

♦ Заметим, что в работе [343] предлагается и ускоренный адаптивный вариант градиентного метода в условиях относительной сильной выпуклости и гладкости. Как известно, в общем случае ускорения такой метод может не давать [246]. Однако в данном контексте удаётся наблюдать ускорение.

Также отметим, что адаптивность ускоренного метода работы [343] хорошо согласуется с общей идей адаптивного подбора параметров, характеризующих гладкость целевой функции, продемонстрированной также в § 5. А именно, из доказательства выделяется условие, следующее из гладкости, и явно зашивается в сам алгоритм. Итерация осуществляется только в том случае, когда неизвестный параметр подобран так, что это условие выполнится. В отличие от параметра, характеризующего сильную выпуклость, в такие условия для параметра гладкости не входят неизвестные величины (кроме самого этого параметра), например само решение задачи x_*.

Несмотря на то, что в упражнениях 3.10 и 3.11 применяется один и тот же метод (3.21), стоит отметить, что применяется он существенно по-разному. В упражнении 3.10 это внутренний метод. В упражнении 3.11, наоборот, это внешний метод — оболочка. Тем не менее для решения вспомогательной задачи, возникающей в упражнении 3.11, можно использовать тензорные методы с шагом, описанным в упражнении 3.10. Тогда получится, что метод (3.21) будет возникать сразу в двух местах. Заметим, что идея применения методов второго порядка (типа Ньютона) к минимизации целевой функции вида суммы может быть мотивирована тем, что вычисление гессиана суммы (если это сумма не очень большого числа слагаемых) может существенно не влиять на сложность итерации, определяющуюся стоимостью обращения (регуляризованного) гессиана. Другими словами, метод как бы не замечает до определённого момента (числа слагаемых), что минимизируется сумма. Время работы метода будет практически такое же, как если бы сумма была из одного слагаемого. Собственно, поэтому описанный в упражнении 3.11 приём, с помощью которого можно сводить задачу оптимизации большого числа слагаемых к задаче минимизации меньшего числа слагаемых такого же типа, даёт дополнительную мотивацию для изучения тензорных методов, см. [25,150,375,473,476] и цитированную в этих работах литературу. ♦

§ 4

Прямодвойственная структура градиентного спуска

Как и в § 2, 3, рассмотрим сначала общую задачу выпуклой оптимизации (2.1):

$$f(x) \to \min_{x \in Q}.$$

Под *прямодвойственным методом* решения задачи (2.1) будем понимать такой метод, сходимость которого может быть сформулирована (представлена) в терминах *сертификата точности* (2.25) [465] (по А. С. Немировскому) или, в общем случае, в терминах неравенств типа (4.2) [32, 75, 76, 150, 477, 481] (по Ю. Е. Нестерову).

В данном параграфе будет продемонстрирована прямодвойственная природа градиентного спуска (2.6) [4]. Сначала на примере задачи минимизации выпуклой функции при аффинных ограничениях демонстрируется, как градиентный спуск применяется к двойственной задаче (решение прямой задачи удаётся восстановить за счёт прямодвойственности метода), а затем (в конце параграфа) градиентный спуск будет применён к исходной (прямой) задаче при дополнительном предположении, что аффинные ограничения имеют достаточно простую структуру (решение двойственной задачи также удаётся восстановить за счёт прямодвойственности метода). Отмеченные возможности прямодвойственных методов, рассмотренные далее на примере только градиентного спуска, отчасти и объясняют их название [481].

♦ В действительности при правильном взгляде [465] практически любой численный метод оптимизации с фиксированными шагами (см. указание к упражнению 1.3 и замечание 1.6) является прямодвойственным. Нетривиальный пример — метод эллипсоидов. Как уже отмечалось ранее, к прямодвойственным методам относят методы, в которых имеются оценки на *сертификат точности* (2.25) [465]. Отметим, что в данном параграфе мы явно не используем сертификат точности, поскольку его использование приводит к наличию дополнительного слагаемого в правой части оценки (2.25), от которого на самом деле можно избавиться. Однако стоит отметить, что в идейном

плане в § 4 используется по сути тот же самый подход, что и в работах [465, 481]. ♦

Пусть сначала планируется решать двойственную задачу. Для двойственных задач множество Q — либо всё пространство, либо неотрицательный ортант, либо прямое произведение пространства на неотрицательный ортант. В любом из этих случаев имеет смысл выбирать 2-норму и евклидову прокс-структуру (см. § 2).

Итак, вернёмся к формуле (2.12) с $h = 1/L$ (см. формулу (2.21)). Перепишем её следующим образом:

$$f(x^{k+1}) \leqslant \{f(x^k) + \langle \nabla f(x^k), x - x^k\rangle\} + \frac{L}{2}\|x - x^k\|_2^2 - \frac{L}{2}\|x - x^{k+1}\|_2^2 + \delta. \quad (4.1)$$

Суммируя неравенства (4.1) по $k = 0, \ldots, N-1$ и учитывая выпуклость функции $f(x)$ и произвол в выборе $x \in Q$, получим

$$f(\bar{x}^N) \leqslant \frac{1}{N} \min_{x \in Q} \left\{ \sum_{k=0}^{N-1} [f(x^k) + \langle \nabla f(x^k),\, x - x^k\rangle] + \frac{L}{2}\|x - x^0\|_2^2 \right\} + \delta, \quad (4.2)$$

где

$$\bar{x}^N = \frac{1}{N} \sum_{k=1}^{N} x^k.$$

Данная формула является обоснованием *прямодвойственности* метода градиентного спуска (2.6), (2.21) [32, гл. 3], [4, 465, 481]. Как будет продемонстрировано ниже, сходимость метода в таком смысле (более сильном, чем просто по функции) позволяет строить сходящуюся с такой же скоростью последовательность и для сопряжённой (двойственной) задачи.

Будем считать, что $\delta \leqslant \varepsilon/2$ (см. формулу (2.18)). Рассмотрим следующий (вычислимый! — ввиду простоты множества Q, см. § 2) критерий останова метода:

$$f(\bar{x}^N) - \frac{1}{N} \min_{x \in B_{R,Q}(x^0)} \left\{ \sum_{k=0}^{N-1} [f(x^k) + \langle \nabla f(x^k),\, x - x^k\rangle] \right\} \leqslant \varepsilon. \quad (4.3)$$

Обратим внимание на то, что минимум в формуле (4.3) берётся по множеству $B_{R,Q}(x^0)$, где $R = \|x_* - x^0\|_2$, а не $B_{R,Q}(x_*)$, поскольку x_* нам неизвестно[1]. При этом в § 2 мы показали, что $x_* \in B_{R,Q}(x^0)$. Зна-

[1] Строго говоря, и $R = \|x_* - x^0\|_2$ также неизвестен. Однако при использовании $B_{R,Q}(x^0)$ удаётся ограничиться лишь одним неизвестным R, по которому можно делать рестарты подобно указанию к упражнению 2.3, а в ряде случаев достаточно здесь исходить из размера множества Q.

чит, в силу выпуклости функции $f(x)$ (см. неравенство (2.13)) имеем нужное нам неравенство

$$f(\bar{x}^N) - f(x_*) \leqslant f(\bar{x}^N) - \frac{1}{N}\sum_{k=0}^{N-1}[f(x^k) + \langle \nabla f(x^k), x_* - x^k\rangle] \leqslant$$
$$\leqslant f(\bar{x}^N) - \frac{1}{N}\min_{x\in B_{R,Q}(x^0)}\left\{\sum_{k=0}^{N-1}[f(x^k) + \langle \nabla f(x^k),\ x - x^k\rangle]\right\} \leqslant \varepsilon. \quad (4.4)$$

С другой стороны, из неравенства (4.2) имеем

$$f(\bar{x}^N) \leqslant \frac{1}{N}\min_{x\in Q}\left\{\sum_{k=0}^{N-1}[f(x^k) + \langle \nabla f(x^k),\ x - x^k\rangle] + \frac{L}{2}\|x - x^0\|_2^2\right\} + \frac{\varepsilon}{2} \leqslant$$
$$\leqslant \frac{1}{N}\min_{x\in B_{R,Q}(x^0)}\left\{\sum_{k=0}^{N-1}[f(x^k) + \langle \nabla f(x^k),\ x - x^k\rangle]\right\} + \frac{LR^2}{2N} + \frac{\varepsilon}{2}. \quad (4.5)$$

Значит, с учётом соотношений (2.5), (2.19) метод (2.6), (2.21) при условии (2.4) гарантированно остановится по критерию (4.3), сделав не более

$$N = \frac{LR^2}{\varepsilon} \leqslant \left(\frac{L_\nu R^{1+\nu}}{\varepsilon}\right)^{2/(1+\nu)} \quad (4.6)$$

итераций (вычислений $\nabla f(x^k)$).

Рассмотрим конкретный пример использования оценки типа (4.2) [4, 234, 270, 584]. Пусть необходимо решить задачу (в данном случае численно решать планируется двойственную задачу к (4.7), поэтому переменные в прямой задаче (4.7) обозначили через y)

$$\varphi(y) \to \min_{Ay=b,\, y\in\widetilde{Q}}, \quad (4.7)$$

где $\varphi(y)$ — μ-сильно выпуклая функция в p-норме на $\widetilde{Q}$ $(1 \leqslant p \leqslant 2)$. Решение задачи (4.7) обозначим через y_*, а оптимальное значение функционала — через φ_* $(\varphi_* = \varphi(y_*))$.

Построим (с точностью до знака) *двойственную задачу* к задаче (4.7):

$$f(x) = \max_{y\in Q}\{\langle x, b - Ay\rangle - \varphi(y)\} \to \min_{x\in\mathbb{R}^n}. \quad (4.8)$$

♦ Опишем общий принцип построения двойственных задач [182, гл. 5]. Итак, пусть исходная задача выпуклой оптимизации имеет вид

$$\varphi(y) \to \min_{h(y)\leqslant 0,\, Ay=b,\, y\in\widetilde{Q}}.$$

Тогда

$$\min_{h(y)\leqslant 0,\, Ay=b,\, y\in\widetilde{Q}} \varphi(y) = \min_{y\in Q}\left\{\varphi(y) + \max_{z\geqslant 0}\langle z, h(y)\rangle + \max_{x}\langle x, Ay-b\rangle\right\} \overset{?}{=}$$

$$\overset{?}{=} \max_{z\geqslant 0,\, x}\min_{y\in Q}\left\{\varphi(y) + \langle z, h(y)\rangle + \langle x, Ay-b\rangle\right\}.$$

Равенство со знаком вопроса обосновывается с помощью теорем типа *фон Неймана* или *Сиона — Какутани* [164, приложение D.4]. К сожалению, при таком подходе требуется компактность множества Q или возможность компактифицировать двойственные переменные (z, x) (см. [477, п. 3.1.8], а также замечание 4.2 и упражнение 4.1). В любом случае в реальных задачах, как правило, удаётся обосновать это равенство [182], которое также называют *сильной двойственностью* [182, гл. 5]. Таким образом, решение исходной задачи сводится к двойственной задаче (с точностью до знака):

$$\max_{y\in Q}\left\{\langle x, b-Ay\rangle - \langle z, h(y)\rangle - \varphi(y)\right\} \to \min_{z\geqslant 0,\, x}. \quad \blacklozenge$$

Точное решение вспомогательной max-задачи (4.8) будем обозначать через $y(x)$. Во многих важных приложениях основной вклад в сложность расчёта $y(x)$ даёт умножение Ay. Это так, например, для сепарабельных функционалов

$$\varphi(y) = \sum_{i=1}^{m}\varphi_i(y_i)$$

и параллелепипедных ограничений $\widetilde{Q}$. В таких случаях задача (4.8) сводится к n задачам одномерной оптимизации, которые с запасом могут быть решены за время (2.2) (см. упражнение 1.4) при условии, что Ay уже было посчитано.

Для двойственного функционала $f(x)$, определяемого согласно соотношению (4.8), выполняется условие (2.4) с $\nu = 1$ и $L_1 = L = \frac{1}{\mu}\max_{\|y\|_p\leqslant 1}\|Ay\|_2^2$ [4, 485]. В частности, для $p = 1$ имеем

$$L = \frac{1}{\mu}\max_{j=1,\ldots,m}\|A^j\|_2^2,$$

где A^j — j-й столбец матрицы A. Для $p = 2$ имеем

$$L = \frac{1}{\mu}\lambda_{\max}(A^TA) \overset{\text{def}}{=} \frac{1}{\mu}\sigma_{\max}(A).$$

Замечание 4.1 (метод регуляризации и техника двойственного сглаживания). Добиться сильной выпуклости функции $\varphi(y)$ всегда можно с помощью *регуляризации* задачи. Опишем, в чём состоит

техника регуляризации (см., например, [8], [14, гл. 9] и цитированную там литературу), восходящая к работам трёх отечественных научных школ: А. Н. Тихонова [94] (Москва), М. М. Лаврентьева [65] (Новосибирск), В. К. Иванова [54] (Свердловск), занимавшихся изучением некорректных задач. Рассмотрим новую задачу:

$$\varphi^{\mu}(y) = \varphi(y) + \mu V(y, y^0) \to \min_{Ay=b,\, y\in\widetilde{Q}}, \tag{4.9}$$

где $V(y, y^0)$ — 1-сильно выпуклая в p-норме функция y. Обозначим через φ^{μ}_* оптимальное значение функционала в задаче (4.9). Пусть[2]

$$\mu \leqslant \frac{\varepsilon}{2V(y_*, y^0)} \tag{4.10}$$

и удалось найти $\varepsilon/2$-решение задачи (4.9), т. е. нашёлся такой вектор $y_{\varepsilon/2}$, что $Ay_{\varepsilon/2} = b$, $y_{\varepsilon/2} \in \widetilde{Q}$ и

$$\varphi^{\mu}(y_{\varepsilon/2}) - \varphi^{\mu}_* \leqslant \frac{\varepsilon}{2}.$$

Тогда

$$\varphi(y_{\varepsilon/2}) - \varphi_* \leqslant \varepsilon.$$

Действительно,

$$\varphi(y_{\varepsilon/2}) - \varphi_* \leqslant \varphi^{\mu}(y_{\varepsilon/2}) - \varphi_* \leqslant \varphi^{\mu}(y_{\varepsilon/2}) - \varphi^{\mu}_* + \frac{\varepsilon}{2} \leqslant \varepsilon.$$

Здесь использовались определение величины φ^{μ}_* и формула (4.10):

$$\varphi^{\mu}_* = \min_{Ay=b,\, y\in\widetilde{Q}} \{\varphi(y) + \mu V(y, y^0)\} \leqslant \varphi(y_*) + \mu V(y_*, y^0) \leqslant \varphi_* + \frac{\varepsilon}{2}.$$

Стоит отметить, что если изначально рассматривалась задача вида (4.8), то говорят, что функционал $f(x)$ представим в *форме Лежандра*. Пусть $\widetilde{Q}$ — выпуклое компактное множество простой структуры. В этом случае описанная выше техника регуляризации $\varphi(y) \to \varphi^{\mu}(y)$, в которой вместо $R^2 = V(y_*, y^0)$ используется $\widetilde{R}^2 = \max_{y\in\widetilde{Q}} V(y, y^0)$ с $\mu \leqslant \varepsilon/(2\widetilde{R}^2)$, приводит к сглаживанию функции:

$$\begin{gathered} f(x) \to f_{\mu}(x) = \max_{y\in\widetilde{Q}} \bigl\{\langle x, b - Ay\rangle - \varphi(y) - \mu V(y, y^0)\bigr\}, \\ 0 \leqslant f(x) - f_{\mu}(x) \leqslant \frac{\varepsilon}{2}. \end{gathered} \tag{4.11}$$

При этом $f_{\mu}(x)$ будет иметь константу Липшица градиента в 2-норме:

$$L_{\varepsilon} = \frac{2\widetilde{R}^2}{\varepsilon} \max_{\|y\|_p \leqslant 1} \|Ay\|_2^2.$$

[2] Как правило, величина $V(y_*, y^0)$ неизвестна, поэтому на практике используются рестарты по параметру μ, приводящие к увеличению общего числа итераций в несколько раз [24, 120], см. также указание к упражнению 2.3.

Простейший пример такого сглаживания:

$$f(x) = \max_{l=1,\dots,m} \langle c^l, x\rangle = \max_{y\in S_m(1)} \sum_{l=1}^m y_l \langle c^l, x\rangle \to$$
$$\to \max_{y\in S_m(1)} \left\{ \sum_{l=1}^m y_l \langle c^l, x\rangle - \mu \sum_{l=1}^m y_l \ln\Big(\frac{y_l}{1/m}\Big)\right\} =$$
$$= \mu \ln\left(\sum_{l=1}^m \exp\Big(\frac{\langle c^l, x\rangle}{\mu}\Big)\right) - \mu \ln m = f_\mu(x),$$

где $\mu = \varepsilon/(2\ln m)$. Описанную выше конструкцию (4.11) обычно называют *двойственным сглаживанием* или *техникой сглаживания по Нестерову* [75, гл. 5], [485]. В классе рассматриваемых в этой главе неускоренных методов данная техника по оценкам не даёт преимуществ: задача выпуклой оптимизации с негладким функционалом для решения с точностью по функции ε требует $\sim \varepsilon^{-2}$ вычислений (суб-)градиента (см. упражнение 2.1), и сглаженная задача также требует $\sim L_\varepsilon \varepsilon^{-1} \sim \varepsilon^{-2}$ вычислений (суб-)градиента. Однако для ускоренных методов техника двойственного сглаживания приводит к лучшим оценкам [75, гл. 5], [477, 485]:

$$\sim \sqrt{L_\varepsilon \varepsilon^{-1}} \sim \sqrt{\varepsilon^{-2}} \sim \varepsilon^{-1}.$$

Разумеется, имеет смысл говорить о двойственном сглаживании только в случае, когда задача максимизации (4.11) является относительно простой. Как следствие, описанная техника сглаживания применима к намного более узкому классу задач, чем регуляризация. Более того, конструкция, описанная в замечании 5.1 в части решения седловых задач, позволяет получать аналогичные результаты при более общих условиях. Тем не менее стоит отметить, что в определённых (композитных) случаях описанная техника позволяет получать новые результаты, недостижимые с помощью техники из замечания 5.1, см., например, [2, 208, 255, 403]. Отметим также, что есть и другие способы сглаживания (см., например, [120]), впрочем, также имеющие весьма ограниченную область применимости.

Хорошо известный пример использования регуляризации — способ вычисления (понимания) *псевдообратной матрицы* [86, 183]:

$$A^+ = \lim_{\mu\to 0+} (A^T A + \mu I)^{-1} A^T.$$

Такое понимание эквивалентно тому, что

$$x_* = A^+ b = \lim_{\mu\to 0+} \arg\min_x \left\{ \frac{1}{2}\|Ax - b\|_2^2 + \frac{\mu}{2}\|x\|_2^2 \right\}$$

является решением задачи

$$\frac{1}{2}\|Ax-b\|_2^2 \to \min_x$$

с наименьшим значением 2-нормы, если решение не единственно, см. также упражнение 5.9. В анализе данных описанная регуляризация имеет простую содержательную интерпретацию — байесовская регуляризация для задачи нормальной регрессии с нормальным (гауссовским) априорным распределением параметров [103, лекции 13, 14]. На рассмотренном примере также удобно демонстрировать связь метода регуляризации и метода *штрафных функций*, см. замечание 4.3 и [135].

Менее известный, но не менее интересный пример *итеративной регуляризации/сглаживания* имеется в работе [300], в которой решение седловой билинейной задачи сводится к последовательности задач выпуклой оптимизации в условиях острого минимума [86, 532]. ■

♦ Так же как и в упражнении 2.3, здесь следует отметить, что *из оптимального метода для сильно выпуклой задачи можно получить с помощью регуляризации оптимальный метод для просто выпуклой задачи*. Во всяком случае, пока не удалось придумать ни одного контрпримера, когда бы это было не так. ♦

Вернёмся к задаче (4.8), в которой для большей наглядности будем считать, что $\widetilde{Q} = \mathbb{R}^{n_y}$, $n_y = \dim y$:

$$f(x) = \max_{y\in\mathbb{R}^{n_y}} \left\{\langle x, b-Ay\rangle - \varphi(y)\right\} \to \min_{x\in\mathbb{R}^n}.$$

Положим[3] $x^0 = 0$, $N = 2LR^2/\varepsilon$, где $R = \|x_*\|_2$. Рассмотрим метод градиентного спуска (1.22):

$$x^{k+1} = x^k - \frac{1}{L}\nabla f(x^k).$$

Подобно неравенству (4.5) можно написать

$$f(\bar{x}^N) \leqslant \frac{1}{N}\min_{x\in\mathbb{R}^n}\left\{\sum_{k=0}^{N-1}[f(x^k) + \langle\nabla f(x^k), x-x^k\rangle] + \frac{L}{2}\|x-x^0\|_2^2\right\} \leqslant$$

$$\leqslant \frac{1}{N}\min_{x\in B_{2R}(0)}\left\{\sum_{k=0}^{N-1}[f(x^k) + \langle\nabla f(x^k), x-x^k\rangle]\right\} + \underbrace{\frac{2LR^2}{N}}_{\varepsilon}.$$

[3] Если выбирать точку старта $x^0 \neq 0$, то оценка R в приводимых далее выкладках (результатах) ухудшится: $R = \|x^0\|_2 + \|x_* - x^0\|_2$. Ухудшатся и числовые множители. Детали см., например, в [354].

Здесь выбирается шар радиуса $2R$, чтобы впоследствии можно было получить оценку (4.14), для чего нужна оценка (4.13) именно с $2R$, что и требует выбора радиуса шара, равного $2R$. Ввиду соотношения (4.8) отсюда *по формуле Демьянова — Данскина* [41, 45, 168]

$$\nabla f(x) = b - Ay(x)$$

получаем, что

$$f(\overline{x}^N) - \frac{1}{N}\sum_{k=0}^{N-1}\langle x^k, b - Ay(x^k)\rangle + \frac{1}{N}\sum_{k=0}^{N-1}\varphi(y(x^k)) - \\ - \frac{1}{N}\min_{x\in B_{2R}(0)}\left\{\sum_{k=0}^{N-1}\langle b - Ay(x^k), x - x^k\rangle\right\} \leqslant \varepsilon. \quad (4.12)$$

В силу выпуклости функции $\varphi(y)$ (см. формулу (4.7)) из неравенства (4.12) следует, что

$$f(\overline{x}^N) + \varphi\underbrace{\left(\frac{1}{N}\sum_{k=0}^{N-1} y(x^k)\right)}_{\bar{y}^N} + \max_{x\in B_{2R}(0)}\left\{\left\langle A\frac{1}{N}\sum_{k=0}^{N-1} y(x^k) - b, x\right\rangle\right\} \leqslant \varepsilon,$$

т. е.

$$f(\overline{x}^N) + \varphi(\overline{y}^N) + 2R\|A\overline{y}^N - b\|_2 \leqslant \varepsilon. \quad (4.13)$$

Из неравенства (4.13) и *слабой двойственности* $-\varphi(y_*) \leqslant f(x_*)$ получаем, что

$$\varphi(\overline{y}^N) - \varphi(y_*) \leqslant \varphi(\overline{y}^N) - \varphi(y_*) + 2R\|A\overline{y}^N - b\|_2 \leqslant \\ \leqslant \varphi(\overline{y}^N) + f(x_*) + 2R\|A\overline{y}^N - b\|_2 \leqslant \\ \leqslant \varphi(\overline{y}^N) + f(\overline{x}^N) + 2R\|A\overline{y}^N - b\|_2 \leqslant \varepsilon.$$

♦ По существу слабая двойственность — это отражение простого факта, что всегда имеет место неравенство

$$\max_y \min_x L(x, y) \leqslant \min_x \max_y L(x, y).$$

На самом деле во всех естественных ситуациях, когда рассматриваются невырожденные (совместные) выпуклые задачи, в этом неравенстве достигается равенство, т. е. имеет место сильная двойственность [182, гл. 5]. ♦

Поскольку x_* одновременно является решением двойственной задачи (4.8) и множителем Лагранжа к ограничению $Ay = b$ в зада-

че (4.7) (см. представление (4.8)), справедливо неравенство

$$f(x_*) = \underbrace{\langle x_*, b - Ay_*\rangle}_{0} - \varphi(y_*) =$$
$$= \max_y \left\{\langle x_*, b - Ay\rangle - \varphi(y)\right\} \geqslant \langle x_*, b - A\overline{y}^N\rangle - \varphi(\overline{y}^N),$$

т. е.

$$\varphi(y_*) - \langle x_*, A\overline{y}^N - b\rangle \leqslant \varphi(\overline{y}^N).$$

Объединяя это неравенство с установленым ранее неравенством

$$\varphi(\overline{y}^N) - \varphi(y_*) + 2R\|A\overline{y}^N - b\|_2 \leqslant \varepsilon,$$

получим

$$R\|A\overline{y}^N - b\|_2 \leqslant \varepsilon.$$

Таким образом, установлен следующий результат.

Теорема 4.1. *Пусть нужно решить задачу* (4.7) *в следующем смысле*:

$$\varphi(\overline{y}^N) - \varphi(y_*) \leqslant \varepsilon, \quad \|A\overline{y}^N - b\|_2 \leqslant \tilde{\varepsilon}. \tag{4.14}$$

Для этого рассмотрим двойственную задачу (4.8), *которую будем решать градиентным спуском* (1.22):

$$x^{k+1} = x^k - \frac{1}{L}\nabla f(x^k)$$

с $x^0 = 0$. *Выберем в качестве критерия останова метода условия* (*зазор двойственности и невязку в ограничении*)

$$f(\overline{x}^N) + \varphi(\overline{y}^N) \leqslant \varepsilon, \quad \|A\overline{y}^N - b\|_2 \leqslant \tilde{\varepsilon},$$

где

$$\overline{x}^N = \frac{1}{N}\sum_{k=1}^{N} x^k, \quad \overline{y}^N = \frac{1}{N}\sum_{k=0}^{N-1} y(x^k),$$

из которых вытекают неравенства (4.14). *Тогда метод гарантированно остановится, сделав не более чем*

$$\max\left\{\frac{2LR^2}{\varepsilon}, \frac{2LR}{\tilde{\varepsilon}}\right\} \tag{4.15}$$

итераций, где

$$L = \frac{1}{\mu}\max_{\|y\|_p \leqslant 1} \|Ay\|_2^2, \quad R = \|x_*\|_2.$$

Если решение x_* *задачи* (4.8) *не единственно, то в оценке* R *в формуле* (4.15) *выбирается то решение, которое имеет наименьшую* 2*-норму.*

Замечание 4.2 (оценка размера решения двойственной задачи). В оценку (4.15) входит неизвестный размер решения двойственной задачи $R = \|x_*\|_2$; см. соотношение (4.8). Если решение x_* не единственно, то выбирается решение, наименьшее по 2-норме (см. §1, 2). Это R можно оценить следующим образом [75, п. 4.3.4], [405]:

$$R^2 = \|x_*\|_2^2 \leqslant \frac{\|\nabla\varphi(y_*)\|_2^2}{\widetilde{\sigma}_{\min}(A)}, \tag{4.16}$$

где

$$\widetilde{\sigma}_{\min}(A) = \min\left\{\lambda > 0 \colon \exists\, x \neq 0 \colon AA^T x = \lambda x\right\}.$$

Действительно, исходя из определения x_* и y_* для любого $y \in \widetilde{Q}$ получаем

$$\begin{aligned} -\varphi(y_*) &= \langle x_*, b - Ay_*\rangle - \varphi(y_*) = \\ &= \max_{y\in Q}\left\{\langle x_*, b - Ay\rangle - \varphi(y)\right\} \geqslant \langle x_*, b - Ay\rangle - \varphi(y) = \\ &= \langle x_*, Ay_* - Ay\rangle - \varphi(y) = \langle A^T x_*, y_* - y\rangle - \varphi(y). \end{aligned}$$

Следовательно, для любого $y \in \widetilde{Q}$ имеем

$$\varphi(y) \geqslant \varphi(y_*) + \langle -A^T x_*, y - y_*\rangle.$$

В силу выпуклости функции $\varphi(y)$ отсюда следует, что

$$-A^T x_* = \nabla\varphi(y_*).$$

Точнее, $-A^T x_* \in \partial\varphi(y_*)$. Осталось только заметить, что для любого $x \in (\operatorname{Ker} A^T)^\perp$ имеет место неравенство

$$\|-A^T x\|_2^2 = \langle -A^T x, -A^T x\rangle = \langle x, AA^T x\rangle \geqslant \widetilde{\sigma}_{\min}(A)\|x\|_2^2. \qquad \blacksquare$$

Пример 4.1 (децентрализованная распределённая оптимизация [262, 403, 405, 460, 539, 540, 590]). Пусть необходимо решать задачу выпуклой оптимизации

$$\sum_{i=1}^n \varphi_i(y) \to \min_{y\in\mathbb{R}}. \tag{4.17}$$

Для большей наглядности считаем y скалярной величиной. Заметим, однако, что от этого упрощения легко отказаться. Будем считать, что $\varphi_i''(y) \geqslant \mu$, $i = 1, \ldots, n$, $y \in \mathbb{R}$. Предположим, что есть связанная сеть (коммуникационный граф) $G = \langle V, E\rangle$ из n узлов. В i-м узле хранится функция $\varphi_i(y)$. Зададим матрицу инцидентности

$$\mathrm{I} = \|\mathrm{I}_{ij}\|_{i,j=1}^n \colon \quad \mathrm{I}_{ij} = 1, \quad (i,j) \in E; \quad \mathrm{I}_{ij} = 0, \quad (i,j) \neq E.$$

По матрице I построим симметричную неотрицательно определённую *матрицу Лапласа* (*Кирхгофа*) $W \succ 0$:

$$W_{ij} = \begin{cases} -\mathrm{I}_{ij}, & i \neq j, \\ \sum\limits_{j=1}^{n} \mathrm{I}_{ij}, & i = j. \end{cases}$$

По *теореме Фробенуса — Перрона* [81, § 7, 8, гл. 2]

$$Wy = 0 \iff y_1 = \ldots = y_n. \tag{4.18}$$

Ввиду соотношения (4.18) перепишем задачу (4.17) следующим образом:

$$\varphi(y) = \sum_{i=1}^{n} \varphi_i(y_i) \to \min_{Wy=0}. \tag{4.19}$$

Построим (с точностью до знака) двойственную задачу к задаче (4.19) (см. формулу (4.8)):

$$f(x) = \varphi^*(-Wx) = \max_{y \in \mathbb{R}^n} \{-\langle x, Wy\rangle - \varphi(y)\} =$$

$$= \sum_{i=1}^{n} \max_{y_i \in \mathbb{R}} \{[-Wx]_i y_i - \varphi_i(y_i)\} \to \min_{x \in \mathbb{R}^n}. \tag{4.20}$$

Считаем, что i-я вспомогательная задача максимизации в формуле (4.20) может эффективно решаться в i-м узле. Заметим, что для функции $f(x)$ константу Липшица градиента можно оценить как $L = \sigma_{\max}(W)/\mu$ [4, 485], а на размер двойственного решения есть оценка (см. замечание 4.2)

$$R^2 = \|x_*\|_2^2 \leqslant \frac{\|\nabla \varphi(y_*)\|_2^2}{\widetilde{\sigma}_{\min}(W)}.$$

Согласно написанному ранее в этом параграфе решать задачу (4.20) можно методом

$$x_i^{k+1} = x_i^k - \frac{1}{L}[\nabla f(x^k)]_i = x_i^k + \frac{1}{L}[W\tilde{y}(-Wx^k)]_i, \tag{4.21}$$

где через $\tilde{y}(-Wx)$ обозначается решение задачи (4.20). Итак, пусть в каждом узле хранятся $\{x_i^k, \tilde{y}_i([-Wx^k]_i)\}_k$. Ключевое наблюдение: чтобы вычислить $\tilde{y}_i([-Wx^k]_i)$, i-му узлу необходимо обратиться только к своим непосредственным соседям за соответствующими компонентами вектора x^k (см. формулу (4.20)), а чтобы вычислить x_i^{k+1}, i-му узлу также необходимо обратиться только к своим непосредственным соседям за соответствующими компонентами вектора $\tilde{y}(-Wx)$ (см. формулу (4.21)). Таким образом, один шаг градиентного спуска

для двойственной задачи приводит к коммуникации каждого узла со своими соседями два раза (передаётся два числа). Поскольку вычислительные возможности узлов, как правило, на несколько порядков выше скорости передачи информации по сети, полученный дисбаланс (решать вспомогательную задачу поиска $\tilde{y}_i([-Wx^k]_i)$ заметно труднее, чем послать и принять несколько чисел) хорошо способствует эффективному решению задачи.

Несложно заметить (см. формулу (4.15)), что время работы алгоритма будет прямо пропорционально числу обусловленности матрицы $W^TW = W^2$, т. е. $\sigma_{\max}(W)/\widetilde{\sigma}_{\min}(W)$. В действительности можно улучшить описанный выше подход, если сделать замену $W \to \sqrt{W}$ [540]:

$$\sqrt{W}y = 0 \;\Leftrightarrow\; y_1 = \ldots = y_n,$$
$$\tilde{y}(-\sqrt{W}x) = \arg\max_{y\in\mathbb{R}^n}\left\{-\langle\sqrt{W}x, y\rangle - \varphi(y)\right\},$$
$$\sqrt{W}x^{k+1} = \sqrt{W}x^k + \frac{1}{L}W\tilde{y}(-\sqrt{W}x^k).$$

Обозначая $z = \sqrt{W}x$, запишем метод в новых переменных[4]:

$$\tilde{y}(z) = \arg\max_{y\in\mathbb{R}^n}\{\langle z, y\rangle - \varphi(y)\}, \quad z^{k+1} = z^k + \frac{1}{L}W\tilde{y}(-z^k).$$

Легко понять, что такой метод также может работать распределённо [405, 460, 540, 590], причём один шаг такого варианта градиентного спуска для двойственной задачи приводит к коммуникации каждого узла со своими соседями всего один раз. Таким образом, можно редуцировать

$$\frac{\sigma_{\max}(W)}{\widetilde{\sigma}_{\min}(W)} \quad \text{к} \quad \frac{\sigma_{\max}(\sqrt{W})}{\widetilde{\sigma}_{\min}(\sqrt{W})} = \sqrt{\frac{\sigma_{\max}(W)}{\widetilde{\sigma}_{\min}(W)}}.$$

На основе ускоренных градиентных методов можно построить более быстрые децентрализованные распределённые алгоритмы решения задачи (4.17), см. [262, 307, 540, 590].

К сожалению, во всех случаях (ускоренном и неускоренном) не удаётся построить адаптивные/универсальные (см. § 5) варианты таких методов (в децентрализованном случае), равно как и не удаётся предложить эффективный (практический) критерий останова методов (см. замечание 2.1 и § 4).

[4] Заметим, что если сделать такую замену в самой двойственной задаче, то придётся вместо условия $z \in \mathbb{R}^n$ писать $z \in \Im\sqrt{W} = (\mathrm{Ker}\,\sqrt{W})^\perp$, что порождает сложности с интерпретацией метода. Описанный в этом примере подход (замена переменных в самом алгоритме) является, пожалуй, наиболее простым способом преодоления этих сложностей.

Отметим также, что между рассмотренной в этом примере задачей и задачами типа распределения ресурсов (см. упражнение 4.7) имеется связь — двойственная задача к задаче о распределении ресурсов имеет вид (4.17). Таким образом, появляется возможность решать задачу распределения ресурсов не централизованным образом, как предлагается в указании к упражнению 4.7, а децентрализованным образом, как в разобранном примере. Детали см., например, в работе [205] (см. также [179, п. 7.3.1]) ■

♦ После работы [179] распределённая оптимизация (Grid-технологии) прочно закрепилась в современном анализе данных, см., например, концепцию Google: Federated Learning [372, 391, 414, 447].

Хорошей практической демонстрацией описанной в примере 4.1 техники является распределённый способ вычисления барицентра Васерштейна, учитывающий наличие явного *представления Лежандра* (сопряжённого представления) для расстояния Монжа — Канторовича — Васерштейна [220, 263, 264, 270, 394, 395, 508, 591], см. также замечание 4.4.

Подобно теореме 1.1 можно распространить градиентный спуск и на невыпуклые задачи распределённой оптимизации [573]. ♦

При описанном выше подходе, к сожалению, возникает невязка в ограничении $Ay = b$ в задаче (4.7). Эту невязку можно полностью устранить, изменив подход. Далее мы будем в основном следовать работе [481]. Немного обобщим постановку задачи (4.7):

$$\varphi(y) = F(y) + g(y) \to \min_{Ay \leqslant b,\, y \in \widetilde{Q}}. \tag{4.22}$$

Вместо входящего в задачу (4.7) равенства $Ay = b$ в задаче (4.22) стали рассматривать неравенство $Ay \leqslant b$ и добавили простой выпуклый композитный член $g(y)$. Будем предполагать, что выпуклая функция $F(y)$ удовлетворяет (только) условию (2.3) (см. также неравенство (2.26)), которое в данном случае будет иметь вид

$$F(y) \leqslant F(z) + \langle \nabla F(z), y - z \rangle + \frac{L}{2}\|y - z\|^2 + \delta.$$

Рассмотрим метод вида (3.19) с шагом $h = 1/L$ (2.21) для задачи (4.22):

$$y^{k+1} = \arg \min_{Ay \leqslant b,\, y \in \widetilde{Q}} \left\{ \langle \nabla F(y^k), y - y^k \rangle + g(y) + LV(y, y^k) \right\}. \tag{4.23}$$

Для наглядности будем считать, что задача (4.23) решается на каждой итерации k явно (точно). В приложениях задача (4.23) может быть простой, например, когда неравенство $Ay \leqslant b$ имеет вид $y \leqslant \overline{y}$ или $y \geqslant \overline{y}$ [32, гл. 1, 3].

Повторяя рассуждения из примера 3.1 (см. формулы (3.12), (3.20)), из соотношения (4.23) подобно оценке (4.2) получим

$$\varphi(\bar{y}^N)\leqslant \min_{Ay\leqslant b, y\in\widetilde{Q}}\left\{\frac{1}{N}\sum_{k=0}^{N-1}\bigl[F(y^k)+\langle\nabla F(y^k),y-y^k\rangle+g(y)\bigr]+\frac{LV(y,y^0)}{N}\right\}+\delta, \tag{4.24}$$

где

$$\bar{y}^N=\frac{1}{N}\sum_{k=1}^{N}y^k.$$

Обозначим множитель Лагранжа к ограничению $Ay\leqslant b$ в неравенстве (4.24) через $\tilde{x}^N\geqslant 0$. Тогда неравенство (4.24) можно переписать следующим образом: для любого $\tilde{y}\in\widetilde{Q}$ имеем

$$\begin{aligned}\varphi(\bar{y}^N)\leqslant \min_{y\in Q}\Biggl\{\frac{1}{N}\sum_{k=0}^{N-1}\underbrace{\bigl[F(y^k)+\langle\nabla F(y^k),y-y^k\rangle+g(y)\bigr]}_{\leqslant\varphi(y)}+\\ +\langle\tilde{x}^N,Ay-b\rangle+\frac{LV(y,y^0)}{N}\Biggr\}+\delta\leqslant\varphi(\tilde{y})+\\ +\langle\tilde{x}^N,A\tilde{y}-b\rangle+\frac{LV(\tilde{y},y^0)}{N}+\delta.\end{aligned} \tag{4.25}$$

Введём, подобно соотношению (4.8), двойственную (с точностью до знака) функцию

$$f(x)=\max_{y\in Q}\bigl\{\langle x,b-Ay\rangle-\varphi(y)\bigr\}. \tag{4.26}$$

Обозначим, как и раньше, через $y(x)$ решение задачи максимизации (4.26). Тогда, если в формуле (4.25) выбрать $\tilde{y}=y(\tilde{x}^N)$ и обозначить $R^2=V(y(\tilde{x}^N),y^0)$, получим

$$0\leqslant\varphi(\bar{y}^N)-\varphi(y_*)\leqslant\varphi(\bar{y}^N)+f(\tilde{x}^N)\leqslant\frac{LR^2}{N}+\delta. \tag{4.27}$$

♦ В отличие от неравенств (4.14), в неравенстве (4.27) используется допустимая точка $\bar{y}^N$: $A\bar{y}^N\leqslant b$, поэтому в нём имеет место оценка снизу $0\leqslant\varphi(\bar{y}^N)-\varphi(y_*)$. Из слабой двойственности имеем

$$\varphi(\bar{y}^N)-\varphi(y_*)+f(\bar{x}^N)-f(x_*)\leqslant\varphi(\bar{y}^N)+f(\bar{x}^N).$$

С учётом этих неравенств из неравенства (4.27) следует, что

$$0\leqslant f(\bar{x}^N)-f(x_*)\leqslant\varphi(\bar{y}^N)+f(\bar{x}^N)\leqslant\frac{LR^2}{N}+\delta.\ \blacklozenge$$

Из формулы (4.27) (с $\delta = \varepsilon/2$) при условии, что функция $F(y)$ удовлетворяет (только) условию (2.27), следует, что метод вида (3.19) гарантированно остановится по критерию

$$\varphi(\bar{y}^N) + f(\bar{x}^N) \leqslant \varepsilon,$$

сделав не более

$$N = \frac{2LR^2}{\varepsilon} \leqslant \Big(\frac{2L_\nu R^{1+\nu}}{\varepsilon}\Big)^{2/(1+\nu)} \tag{4.28}$$

итераций (вычислений $\nabla F(y^k)$).

Описанный способ решения задачи композитной оптимизации (4.22) может быть осуществлён в модельной общности (см. § 3) [99].

В заключение подчеркнём, в чём различие в описанных в этом параграфе подходах. В подходе (4.23) решается исходная задача (4.22), в то время как в подходе, описанном в первой половине параграфа, решается двойственная задача (4.8). Отметим, что оба описанных подхода можно сочетать друг с другом [18].

Упражнение 4.1 (условие Слейтера). Рассматривается задача выпуклой оптимизации

$$f(x) \to \min_{h(x)\leqslant 0,\, x\in Q},$$

где $h\colon \mathbb{R}^n \to \mathbb{R}^m$. Двойственная задача (с точностью до знака) имеет вид

$$\varphi(y) = \max_{x\in Q}\big\{-\langle y, h(x)\rangle - f(x)\big\} \to \min_{y\in\mathbb{R}^m_+}. \tag{4.29}$$

Обозначим через y^* решение двойственной задачи. Предположим, что выполняется *условие Слейтера*:

существует такая точка $\bar{x} \in Q$, что $h(\bar{x}) < 0$.

Пусть $\gamma = \min_{i=1,\dots,m}\{-h_i(\bar{x})\}$. Покажите, что

$$\|y^*\|_1 \leqslant \frac{1}{\gamma}(f(\bar{x}) + \varphi(0)) = \frac{1}{\gamma}\big(f(\bar{x}) - \min_{x\in Q} f(x)\big).$$

Указание. См. [169]. Ключевое неравенство:

$$\varphi(0) \geqslant \varphi(y^*) = \max_{x\in Q}\Big\{-\sum_{i=1}^m y_i^* h_i(x) - f(x)\Big\} \geqslant -\sum_{i=1}^m y_i^* h_i(\bar{x}) - f(\bar{x}).$$

Аналогичным образом можно получать оценки на размер решения двойственной задачи и в более общих случаях (см., например, [27]).

Упражнение 4.2. Пусть задача из упражнения 4.1 решена в следующем смысле: найден такой вектор $\tilde{y} \in \mathbb{R}^m_+$, что

$$\langle \tilde{y}, \nabla\varphi(\tilde{y})\rangle \leqslant \varepsilon, \quad \left\|[-\nabla\varphi(\tilde{y})]_+\right\|_2 \leqslant \tilde{\varepsilon},$$

где в соответствии с формулой Демьянова — Данскина [182, гл. 3] $\nabla\varphi(\tilde{y}) = -h(x(\tilde{y}))$, а функция $x(\tilde{y})$ — решение вспомогательной задачи максимизации (4.29). Тогда

$$f(x(\tilde{y})) - f(x_*) \leqslant \varepsilon, \quad \left\|[h(x(\tilde{y}))]_+\right\|_2 \leqslant \tilde{\varepsilon}.$$

Указание. Ключевая выкладка:

$$-\langle \tilde{y}, h(x(\tilde{y}))\rangle - f(x(\tilde{y})) \geqslant -\langle \underbrace{\tilde{y}}_{\tilde{y}\geqslant 0}, \underbrace{h(x_*)}_{h(x_*)\leqslant 0}\rangle - f(x_*) \geqslant -f(x_*).$$

Заметим также, что если бы здесь вместо ограничения в виде неравенства $h(x) \leqslant 0$ мы имели аффинное ограничение в виде равенства $Ax - b = 0$, то в проведённых рассуждениях нужно было бы сделать следующую корректировку: $\nabla\varphi(\tilde{y}) = b - Ax(\tilde{y})$, и, как следствие, условие $\|\nabla\varphi(\tilde{y})\|_2 \leqslant \tilde{\varepsilon}$ обеспечивает выполнение условия

$$\|Ax(\tilde{y}) - b\|_2 \leqslant \tilde{\varepsilon}.$$

Упражнение 4.3. 1. Рассматривается задача поиска *седловой точки* вида (4.8)

$$f(x) = \max_{y\in Q}\left\{\langle x, b - Ay\rangle - \varphi(y)\right\} \to \min_{x\in\mathbb{R}^n},$$

где функция $\varphi(y)$ является μ-сильно выпуклой относительно p-нормы ($1 \leqslant p \leqslant 2$). Покажите, что функция $f(x)$ будет гладкой с константой Липшица градиента в 2-норме:

$$L = \frac{1}{\mu}\max_{\|z\|_p\leqslant 1}\|Az\|_2^2.$$

Более того, если $y_\delta(x)$ — решение вспомогательной задачи максимизации с точностью по функции δ, то

$$\left(\langle x, b - Ay_\delta(x)\rangle - \varphi(y_\delta(x)); b - Ay_\delta(x)\right)$$

будет $(\delta, 2L)$-моделью функции $f(x)$ в точке x относительно 2-нормы (см. начало § 3).

Предложите численный метод решения поставленной (седловой) задачи и сравните его трудоёмкость с нижними оценками [502]. Если $\varphi(y)$ дополнительно имеет липшицев градиент, можно ли это как-то использовать (см. текст, следующий после упражнения 4.8 до замечания 4.4 включительно)?

Покажите, что если вместо $\langle x, b - Ay\rangle$ в определении $f(x)$ использовать функцию $F(x, y)$, выпуклую по x, вогнутую по y, а также достаточно гладкую:

$$\left\|\nabla_x F(x_2, y) - \nabla_x F(x_1, y)\right\|_2 \leqslant L_{xx}\|x_2 - x_1\|_2,$$
$$\left\|\nabla_x F(x, y_2) - \nabla_x F(x, y_1)\right\|_2 \leqslant L_{xy}\|y_2 - y_1\|_2,$$

то

$$\big(F(x, y_\delta(x)) - \varphi(y_\delta(x)); \nabla_x F(x, y_\delta(x))\big)$$

будет $(2\delta, 2L)$-моделью функции $f(x)$ в точке x относительно 2-нормы, где

$$L = L_{xx} + \frac{2L_{xy}^2}{\mu}.$$

2. Задачу вида (4.7) можно решать с помощью *модифицированной функции Лагранжа* (*augmented Lagrangians*) [167]. В основе подхода лежит идея переписывания задачи (4.7) следующим образом ($\mu \geqslant 0$ — выбираемый параметр):

$$\varphi(y) + \frac{\mu}{2}\|Ay - b\|_2^2 \to \min_{Ay = b,\, y \in \widetilde{Q}}$$

и стандартный переход к двойственной задаче:

$$f(x) = \max_{y \in Q} \underbrace{\left\{\langle x, b - Ay\rangle - \varphi(y) - \frac{\mu}{2}\|Ay - b\|_2^2\right\}}_{\Psi(y,x)} \to \min_{x \in \mathbb{R}^n}.$$

Покажите, что если $y_\delta(x)$ — решение вспомогательной задачи максимизации в смысле (3.5):

$$\max_{y \in Q}\left\langle \nabla_y \Psi(y_\delta(x), x), y - y_\delta(x)\right\rangle \leqslant \delta,$$

то

$$\left(\langle x, b - Ay_\delta(x)\rangle - \varphi(y_\delta(x)) - \frac{\mu}{2}\|Ay_\delta(x) - b\|_2^2; b - Ay_\delta(x)\right)$$

будет (δ, μ^{-1})-моделью функции $f(x)$ в точке x относительно 2-нормы (см. начало § 3).

Указание. См. [2, 26, 208, 235, 344, 485, 596, 616].

♦ Описанная в первом пункте упражнения 4.3 конструкция позволяет, в частности, решать гладкие μ_x-сильно выпуклые и μ_y-сильно вогнутые седловые задачи (см. замечание 5.1) за[5] $\widetilde{O}(1/\sqrt{\mu_x\mu_y})$ вы-

[5] Здесь и далее в этом абзаце для наглядности в оценках приводится зависимость только от μ_x, μ_y и ε.

числений ∇_x и $\widetilde{O}(1/\sqrt{\mu_x^2\mu_y})$ вычислений ∇_y или, наоборот (в зависимости от того, что нам выгоднее), за $\widetilde{O}(1/\sqrt{\mu_x\mu_y})$ вычислений ∇_y и $\widetilde{O}(1/\sqrt{\mu_x\mu_y^2})$ вычислений ∇_x, что улучшает наилучшую известную сейчас оценку $\widetilde{O}(1/\min\{\mu_x,\mu_y\})$ (нижняя оценка[6] — $\widetilde{O}(1/\sqrt{\mu_x\mu_y})$ [616]) на число вычислений ∇_x и ∇_y (см. указание к упражнению 5.4), например, по числу вычислений ∇_x за счёт ухудшения оценки на число вычислений ∇_y. Если же, скажем, по x есть только выпуклость, то за счёт регуляризации можно обеспечить, чтобы выполнялась оценка $\mu_x \simeq \varepsilon/R^2$, см. замечание 4.1. В этом случае седловую задачу можно решить за $\widetilde{O}(1/\sqrt{\mu_y\varepsilon})$ вычислений ∇_x и $\widetilde{O}(1/\sqrt{\mu_y^2\varepsilon})$ вычислений ∇_y. Отметим также, что для седловых задач вида суммы [344] и ускоренных рандомизированных алгоритмов из замечания 1 приложения приведённые здесь оценки можно дополнительно улучшить. Также приведённые оценки можно дополнительно улучшить, если одна из размерностей $\dim x$ или $\dim y$ мала [301].

Описанный во втором пункте упражнения 4.3 метод модифицированной функции Лагранжа лежит в основе одного из самых популярных алгоритмов распределённой оптимизации ADMM [179, 403, 501]. ♦

Замечание 4.3 (метод штрафных функций). Метод модифицированной функции Лагранжа тесно связан с методом *штрафных функций* [13, § 16, гл. 5]. Для полноты картины приведём здесь соответствующие идеи. Вместо исходной, вообще говоря, невыпуклой задачи условной оптимизации

$$f(x) \to \min_{g(x)=0} \tag{4.30}$$

рассматривается задача безусловной оптимизации:

$$f(x) + \frac{K}{2}\|g(x)\|_2^2 \to \min_x. \tag{4.31}$$

[6] В работе [435] было показано, как каталист (см. также замечание 3.3 и приложение) позволяет получить (с точностью до логарифмических множителей) оптимальную оценку на базе описанного подхода. Исходная седловая задача переписывается как задача оптимизации $f(x) = \max_y F(x,y) \to \min_x$. Эту задачу оптимизации предлагается решать ускоренным проксимальным методом (см. замечание 3.3), на каждой итерации которого необходимо решать задачу вида $\min_x \max_y\{F(x,y) + \|x - x^k\|_2^2\}$. Число таких итераций будет составлять $\widetilde{O}(1/\sqrt{\mu_x})$. Вспомогательная седловая задача на каждой итерации может быть решена за $\widetilde{O}(1/\sqrt{\mu_y})$ вычислений ∇_x и ∇_y. Таким образом, в итоге получается желаемая оценка [26, 596].

К задаче (4.31) можно прийти, например, *релаксировав* исходную постановку следующим образом [135]:

$$f(x) \to \min_{\frac{1}{2}\|g(x)\|_2^2 \leqslant \frac{1}{2}\varepsilon^2}.$$

В таком случае $K := K(\varepsilon)$ можно понимать как множитель Лагранжа к ограничению

$$\frac{1}{2}\|g(x)\|_2^2 \leqslant \frac{1}{2}\varepsilon^2.$$

Обозначим решение задачи (4.31) через x^K, а решение исходной задачи (4.30) через x_*. Тогда также имеет место следующая связь метода множителей Лагранжа и метода штрафных функций (см., например, [13, § 17, гл. 5], [86, п. 5, § 2, гл. 8]):

$$Kg(x^K) \xrightarrow[K\to\infty]{} \lambda, \quad \text{т. е.} \quad g(x^K) \simeq \frac{\lambda}{K},$$

$$f(x^K) - f(x_*) = \mathrm{O}\left(\frac{\|\lambda\|_2^2}{K}\right),$$

где λ — множитель Лагранжа к ограничению $g(x) = 0$. Метод модифицированной функции Лагранжа является промежуточным методом между отмеченными двумя и может быть проинтерпретирован как их комбинация (сочетание). Метод штрафных функций является одним из наиболее простых и универсальных способов сведения задач условной оптимизации к задачам безусловной оптимизации [38].

Для задач выпуклой оптимизации с ограничениями (см. упражнение 4.1)

$$f(x) \to \min_{h(x)\leqslant 0,\, x\in Q},$$

где $h\colon \mathbb{R}^n \to \mathbb{R}^m$, в случае возможности ограничить решение двойственной задачи $\lambda_* \in Q_\lambda \subseteq \mathbb{R}^m_+$ (например, из соображений упражнения 4.1) можно построить выпуклую негладкую точную штрафную функцию [477, п. 3.1.7]:

$$x_* \in \operatorname{Arg}\min_{h(x)\leqslant 0, x\in Q} f(x) = \operatorname{Arg}\min_{x\in Q}\max_{\lambda\in\mathbb{R}^m_+}\left\{f(x) + \langle\lambda, h(x)\rangle\right\} =$$

$$= \operatorname{Arg}\min_{x\in Q}\left\{f(x) + \max_{\lambda\in\mathbb{R}^m_+}\langle\lambda, h(x)\rangle\right\} = \operatorname{Arg}\min_{x\in Q}\left\{f(x) + \max_{\lambda\in Q_\lambda}\langle\lambda, h(x)\rangle\right\}. \quad \blacksquare$$

♦ На метод штрафных функций (с немного другой функцией штрафа — см. формулу (1.43)) в случае аффинных ограничений $g(x) = Ax - b = 0$ можно посмотреть и с точки зрения двойственного сглаживания, см. замечание 4.1. Действительно, задачу

$$f(x) \to \min_{Ax=b}$$

можно переписать следующим образом:

$$f(x) + \sup_{\lambda} \langle \lambda, Ax - b \rangle \to \min_{x}.$$

Существует такой λ_* (множитель Лагранжа), что последняя задача равносильна задаче

$$f(x) + \langle \lambda_*, Ax - b \rangle \to \min_{x}.$$

Будем считать, что $\|\lambda_*\|_2 \leqslant R_\lambda$. Тогда исходная задача равносильна следующей:

$$f(x) + \max_{\|\lambda\|_2 \leqslant R_\lambda} \langle \lambda, Ax - b \rangle \to \min_{x}.$$

Отсюда заключаем, что $\varepsilon/2$-решение задачи

$$f(x) + \max_{\|\lambda\|_2 \leqslant R_\lambda} \left\{ \langle \lambda, Ax - b \rangle - \frac{\varepsilon}{2R_\lambda^2} \|\lambda\|_2^2 \right\} =$$

$$= f(x) + \begin{cases} \dfrac{R_\lambda^2}{2\varepsilon} \|Ax - b\|_2^2, & \|Ax - b\|_2 \leqslant \dfrac{\varepsilon}{R_\lambda}, \\ R_\lambda \|Ax - b\|_2 - \dfrac{\varepsilon}{2}, & \|Ax - b\|_2 > \dfrac{\varepsilon}{R_\lambda} \end{cases} \to \min_{x}$$

будет ε-решением исходной.

Близкий способ формирования штрафа [135][7] (следует сравнить с задачей (4.31) и выбором $K(\varepsilon)$)

$$F_\varepsilon(x) = f(x) + \frac{R_\lambda^2}{\varepsilon} \|Ax - b\|_2^2 \to \min_{x}$$

---

[7] Данный критерий также можно понимать как (байесовскую) свёртку двух критериев $f(x) \to \min_x$, $\|Ax - b\|_2^2 \to \min_x$.

Чтобы определить, с какими весами сворачивать эти критерии (не ограничивая общности, достаточно из двух весов оставить только вес при квадратичном функционале, полагая вес при $f(x)$ равным 1), нужно вспомнить, что мы хотим добиться выполнения неравенств (см. соотношения (4.14))

$$f(x^N) - f(x_*) \leqslant \varepsilon, \quad \|Ax^N - b\|_2 \leqslant \frac{\varepsilon}{R_\lambda}.$$

Переписывая последнее условие как

$$\frac{R_\lambda^2}{\varepsilon} \|Ax^N - b\|_2^2 \leqslant \varepsilon,$$

получим, что критерии

$$f(x) \to \min_{x}, \quad \frac{R_\lambda^2}{\varepsilon} \|Ax - b\|_2^2 \to \min_{x}$$

следует сворачивать с одинаковыми весами (равными 1), что и было сделано.

также позволяет восстанавливать решение исходной задачи. Действительно, из неравенства

$$F_\varepsilon(x^N) - \min_x F_\varepsilon(x) \leqslant \varepsilon,$$

следует, что

$$f(x^N) - f(x_*) + \frac{R_\lambda^2}{\varepsilon}\|Ax^N - b\|_2^2 \leqslant \varepsilon.$$

В частности, $f(x^N) - f(x_*) \leqslant \varepsilon$. Ввиду неравенства (см. вывод теоремы 4.1 в обозначениях $f \to \varphi$, $x \to y$, $y \to -\lambda$)

$$-R_\lambda\|Ax - b\|_2 \leqslant \langle \lambda_*, Ax - b\rangle \leqslant f(x) - f(x_*)$$

отсюда имеем

$$-R_\lambda\|Ax^N - b\|_2 + \frac{R_\lambda^2}{\varepsilon}\|Ax^N - b\|_2^2 \leqslant \varepsilon.$$

Из этого следует, что

$$R_\lambda\|Ax^N - b\|_2 \leqslant \frac{1+\sqrt{5}}{2}\varepsilon < 2\varepsilon.$$

Заметим, что если $f(x)$ — негладкая выпуклая функция с константой Липшица M, то с помощью слайдинга Дж. Лана [404] можно решить исходную задачу с указанной выше точностью

$$F_\varepsilon(x^N) - \min_x F_\varepsilon(x) \leqslant \varepsilon,$$

используя $O(M^2R_x^2/\varepsilon^2)$ вычислений $\nabla f(x)$ и

$$O\left(\sqrt{\frac{\lambda_{\max}(A^TA)R_\lambda^2R_x^2}{\varepsilon^2}}\right) = O\left(\frac{\lambda_{\max}(\sqrt{A^TA})R_\lambda R_x}{\varepsilon}\right)$$

умножений A^TAx. Этот результат при $A = \sqrt{W}$ и $b = 0$ (см. пример 4.1) позволяет довольно просто объяснить различие между числом коммуникационных шагов и числом вызовов оракула (выдающего $\nabla f_k(x)$) в задачах негладкой децентрализованной распределённой оптимизации, а также построить оптимальные методы децентрализованной распределённой оптимизации для гладких задач, см. [262, 405] и замечание 4.4. ♦

Упражнение 4.4. Технику регуляризации, описанную в замечании 4.1, можно применять не только к задаче (4.7), но и к задаче (4.8):

$$f^\mu(x) = f(x) + \frac{\mu}{2}\|x\|_2^2 \to \min_{x\in\mathbb{R}^n}. \tag{4.32}$$

Обозначим через x_*^μ решение задачи (4.32). Покажите, что

$$\|\nabla f(x)\|_2 \leqslant \|\nabla f^\mu(x)\|_2 + \mu\|x\|_2,$$
$$\langle x, \nabla f(x)\rangle \leqslant \frac{L_\mu}{\mu}(f^\mu(x) - f^\mu(x_*^\mu)), \quad L_\mu = L + \mu,$$
$$\|x_*^\mu\|_2^2 \leqslant \frac{2}{\mu}(f(0) - f(x_*)),$$

где для всех $x, y \in \mathbb{R}^n$ по постановке задачи имеет место неравенство

$$\|\nabla f(y) - \nabla f(x)\|_2 \leqslant L\|y - x\|_2.$$

Используя эти оценки и упражнение 4.2, исследуйте скорость сходимости метода (1.3) с шагом $h = 1/L_\mu$ и $x^0 = 0$, сходящегося по оценке (1.24) на задаче (4.32) с критерием останова

$$\langle x^N, \nabla f(x^N)\rangle \quad \left(\leqslant \frac{L_\mu}{\mu}(f^\mu(x^N) - f^\mu(x_*^\mu))\right) \leqslant \varepsilon,$$
$$\|\nabla f(x^N)\|_2 \qquad \left(\leqslant \|\nabla f^\mu(x^N)\|_2 + \mu\|x^N\|_2\right) \leqslant \tilde{\varepsilon}.$$

Учитывая, что $\|x^N\|_2 \leqslant 2\|x_*^\mu\|_2$ (см. неравенство (1.12)), предложите способ выбора параметра регуляризации μ. Сопоставьте полученную таким образом оценку скорости сходимости с оценкой (4.16), учитывая, что $\|x_*^\mu\|_2 \leqslant \|x_*\|_2$.

Обобщите полученные результаты на случай, когда в исходной и регуляризованной постановке задачи (4.32) вместо $x \in \mathbb{R}^n$ стоит $x \in \mathbb{R}^{n_1} \otimes \mathbb{R}_+^{n_2}$.

Указание. Детали см., например, в работе [24]. Из неравенства (1.8) имеем

$$f^\mu(x) - f^\mu(x_*^\mu) \geqslant \frac{\|\nabla f^\mu(x)\|_2^2}{2L_\mu} = \frac{\|\nabla f(x) + \mu x\|_2^2}{2L_\mu} \geqslant \frac{\mu\langle \nabla f(x), x\rangle}{L_\mu}.$$

Из неравенства (1.14) имеем

$$\frac{\mu}{2}\|x_\mu^*\|_2^2 = \frac{\mu}{2}\|0 - x_\mu^*\|_2^2 \leqslant f^\mu(0) - f^\mu(x_*^\mu) \leqslant f(0) - f(x_*).$$

Последнее неравенство имеет место ввиду соотношений $f^\mu(0) = f(0)$ и

$$f^\mu(x_*^\mu) = f(x_*^\mu) + \frac{\mu}{2}\|x_*^\mu\|_2^2 \geqslant f(x_*) + \frac{\mu}{2}\|x_*^\mu\|_2^2 \geqslant f(x_*).$$

В случае наличия у $f(x)$ представления (4.8) имеет место также оценка

$$f(0) - f(x_*) \leqslant \min_{Ay=b,\, y\in\widetilde{Q}} \varphi(y) - \min_{y\in Q}\varphi(y).$$

Заметим, что для оценки $\|x_*^\mu\|_2$ сверху можно было бы также использовать неравенство $\|x_*^\mu\|_2 \leqslant \|x_*\|_2$ и оценку (4.16).

Упражнение 4.5. Обобщите рассуждения из § 4 на случай, когда в задаче (4.22) вместо неравенства $Ay \leqslant b$ рассматриваются общие выпуклые ограничения: $Ay = b$, $h(y) \leqslant 0$.

Указание. См. [481].

Упражнение 4.6 (сложность проектирования). В самом начале § 2 была приведена оценка (2.2). Покажите, что если множество Q есть шар в p-норме и(или) задаётся небольшим числом сепарабельных выпуклых неравенств вида $\sum_{i=1}^n h_i^j(x_i) \leqslant 0$, $j = 1, \ldots, m$, то задачи вида (2.6), (2.29) и при определённых условиях (3.3) могут быть решены (в смысле (3.4)) за время $O(nm^2 \ln^2(n/\varepsilon))$.

Указание. Характерный пример получения такой оценки разбирается в работе [27] (см. также [164, п. 5.3.3]). В основе подхода — решение малоразмерной двойственной задачи каким-нибудь прямодвойственным быстро (линейно) сходящимся методом типа метода эллипсоидов [219, 465], см. также упражнения 1.4, 5.5. Предварительно двойственные переменные компактифицируются (см. упражнение 4.1), а на заключительном этапе при рассмотрении исходной (прямой) задачи уже используется оценка из упражнения 3.1. Для расчёта градиента двойственного функционала необходимо решить n одномерных задач с точным оракулом, что может быть сделано за линейное время (см. упражнение 1.4), и, поскольку двойственную задачу мы также можем решать за линейное время, все «огрубления», накопленные по ходу описанных рассуждений, соберутся под знаком логарифма и испортят лишь мультипликативную константу в итоговой оценке. Отметим также, что в формуле (2.29) в функционале присутствует не сепарабельное слагаемое вида (см. табл. 1 в § 2):

$$\|x\|_p^2 = \left(\sum_{i=1}^n |x_i|^p \right)^{2/p}.$$

Однако, введя новую переменную $y = \|x\|_p^2$, можно занести это слагаемое в ограничение, заменив в функционале $\|x\|_p^2$ на y и добавив сепарабельное выпуклое ограничение вида неравенства $\|x\|_p^p \leqslant y^{p/2}$, где $p/2 < 1$.

Упражнение 4.7 («нащупывание» цен по Вальрасу и централизованная распределённая оптимизация [17, 55, 286, 353, 354]). Пусть руководство города владеет n пекарнями. Затраты i-й пекарни на выпечку x_i тонн хлеба в день равны $f_i(x_i)$; это сильно выпуклые возрастающие функции. Задача руководства — производить не мень-

ше C тонн хлеба в день (C — объём спроса на хлеб в день со стороны населения города) так, чтобы суммарные затраты всех пекарен были минимальны. Формально задача может быть поставлена следующим образом:

$$\sum_{i=1}^{n} f_i(x_i) \to \min_{\substack{\sum_{i=1}^{n} x_i \geqslant C \\ x_i \geqslant 0,\ i=1,\ldots,n}}. \tag{4.33}$$

Обозначим решение этой задачи $x^* = \{x_i^*\}_{i=1}^n$.

1. Предположим теперь, что у пекарен есть собственники, которые продают хлеб руководству города (распределяющему этот хлеб среди населения) по цене p^k в k-й день. Таким образом, собственники решают задачи

$$x_i(p^k) = \arg\max_{x_i \geqslant 0} \Big\{ \overbrace{\underbrace{p^k x_i}_{\text{выручка}} - \underbrace{f_i(x_i)}_{\text{затраты}}}^{\text{прибыль}} \Big\}, \quad i = 1, \ldots, n. \tag{4.34}$$

Руководство города действует по следующему правилу: каждый день k у него есть представление о том, в каком отрезке лежит равновесная цена $[p^k_{\min}, p^k_{\max}]$. Выставив цену

$$p^k = \frac{1}{2}(p^k_{\min} + p^k_{\max}),$$

руководство собирает следующую информацию с пекарен: $\sum_{i=1}^n x_i(p^k)$. Далее,

$$[p^{k+1}_{\min}, p^{k+1}_{\max}] = \Big[p^k_{\min}, \frac{1}{2}(p^k_{\min} + p^k_{\max})\Big], \quad \text{если} \quad \sum_{i=1}^{n} x_i(p^k) > C,$$

$$[p^{k+1}_{\min}, p^{k+1}_{\max}] = \Big[\frac{1}{2}(p^k_{\min} + p^k_{\max}), p^k_{\max}\Big], \quad \text{если} \quad \sum_{i=1}^{n} x_i(p^k) \leqslant C.$$

Покажите, что

$$x_i(p^k) \xrightarrow[k\to\infty]{} x_i^*.$$

Попробуйте оценить скорость сходимости. Предложите способ оценки $[p^0_{\min}, p^0_{\max}]$.

2. Переписав задачу (4.33) следующим (равносильным) образом:

$$\sum_{i=1}^{n} f_i(x_i) \to \min_{\substack{x_i \geqslant y_i,\ i=1,\ldots,n \\ \sum_{i=1}^{n} y_i \geqslant C \\ x_i, y_i \geqslant 0,\ i=1,\ldots,n}},$$

попробуйте предложить алгоритм нащупывания равновесной цены, когда каждый день пекарни выставляют свою цену на хлеб, по которой

готовы (хотят) продавать хлеб руководству, а руководство закупает хлеб у пекарни, выставившей самую низкую цену. Если таких пекарен, выставивших наименьшую (одинаковую) цену, несколько, то руководство города каким-то (произвольным) образом может распределять закупки среди таких пекарен (и только таких — даже если суммарно эти пекарни производят меньше хлеба, чем нужно руководству).

Указание. Заметим, что пекарни не имеют информации о производственных процессах друг друга, а руководство не имеет представления о производственных процессах на всех пекарнях. С одной стороны, описанный выше процесс можно понимать как процесс нащупывания равновесной цены [85, гл. 10]. С другой стороны, этот процесс можно понимать как распределённый централизованный алгоритм решения задачи выпуклой оптимизации (4.33) [460]: задача хранится на n рабочих узлах/пекарнях, взаимодействие которых осуществляется через центр/руководство, см. также рис. 10 (с точностью до $n-1 \to n$). На каждом узле осуществляется работа только со своей частью задачи (решаются вспомогательные задачи (4.34)), коммуникация осуществляется, как показано на рис. 9.

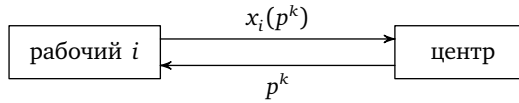

Рис. 9

♦ Централизованная распределённая оптимизация очень похожа на параллельные вычисления. Однако принципиально важным отличием от параллельных вычислений является наличие распределённой (оперативной) памяти, в которой можно хранить описание большой задачи, не пытаясь собрать всю задачу целиком где-то в одном месте [170]. ♦

1. Для решения задачи нужно построить двойственную задачу (с точностью до знака):

$$\psi(p) = \sum_{i=1}^{n} \big(px_i(p) - f_i(x_i(p))\big) - Cp \to \min_{p \geqslant 0}$$

и заметить, что

$$\psi'(p) \overset{\text{def}}{=} \frac{d\psi(p)}{dp} = \sum_{i=1}^{n} x_i(p) - C.$$

Из условий задачи $p^0_{\min} \geqslant 0$, а $p^0_{\max}$ можно оценить с помощью упражнения 4.1. Далее для решения двойственной задачи можно использовать метод деления отрезка пополам, см. упражнение 1.4.

2. В данном случае двойственная задача (с точностью до знака) будет иметь вид

$$\psi(p) := \psi(p_1, \ldots, p_n) = \\ = \sum_{i=1}^{n} \bigl(p_i x_i(p_i) - f_i(x_i(p_i))\bigr) - C \min_{i=1,\ldots,n} p_i \to \min_{p=(p_1,\ldots,p_n)\in\mathbb{R}^n_+}, \quad (4.35)$$

где p_i — множитель Лагранжа к ограничению $x_i \geqslant y_i$. Тогда

$$\partial\psi(p) = \begin{pmatrix} x_1(p) \\ \vdots \\ x_n(p) \end{pmatrix} - C \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix}, \quad \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix} \in S_n(1), \quad \lambda_i > 0 \;\Rightarrow\; i \in \operatorname{Arg}\min_{l=1,\ldots,n} p_l.$$

Двойственная задача (4.35) получилась негладкой, потому что функционал исходной (прямой) задачи не зависел от $\{y_i\}_{i=1}^n$, т. е. не был сильно выпуклым по всей совокупности прямых переменных, см. также указание к упражнению 4.8. Для решения задачи (4.35) можно использовать, например, субградиентный метод [17, 481] или *сходящийся субградиентный метод* Нестерова — Шихмана, который будет иметь в данном контексте вполне естественную интерпретацию [482, 483]. Отличие этого метода от близкого ему субградиентного метода [17], описанного также в упражнении 2.1, в том, что сходимость по функции теперь будет иметь место в обычном (не чезаровском) смысле, поэтому в название метода и вошло слово *сходящийся*. Подумайте, можно ли ускорить процедуры нащупывания равновесия [17, 354, 482, 483], сохранив возможность содержательной интерпретации, если смотреть на слагаемое $C \cdot \min_{i=1,\ldots,n} p_i$ в задаче (4.35) как на композитный член (см. пример 3.1).

♦ В развитие примера 4.1 и упражнения 4.7 отметим, что для так называемой *звёздной топологии* коммуникационного графа (рис. 10), в которой выделяется главный узел (master node), связанный со всеми остальными $n-1$ узлами (slave nodes), диаметр графа равен 2, а $\sigma_{\max}(W)/\widetilde{\sigma}_{\min}(W) = n^2$.

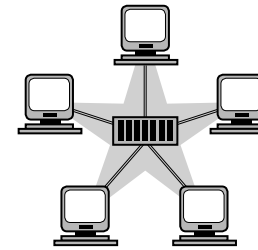

Рис. 10

При этом если (симметричную) матрицу W, которая в данном случае соответствует си-

стеме линейных уравнений

$$\begin{aligned}(n-1)y_1 - y_2 - \ldots - y_n &= 0,\\ -y_1 + y_2 &= 0,\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ -y_1 + y_n &= 0,\end{aligned}$$

определять из немного другой системы

$$\begin{aligned}\frac{n-1}{n}y_1 - \frac{1}{n}y_2 - \ldots - \frac{1}{n}y_n &= 0,\\ -y_1 + y_2 &= 0,\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ -y_1 + y_n &= 0,\end{aligned}$$

то для получившейся в результате (несимметричной) матрицы W выполняется соотношение

$$\frac{\sigma_{\max}(W)}{\widetilde{\sigma}_{\min}(W)} \approx n.$$

Это утверждение следует из того, что спектр новой матрицы W будет иметь вид $[0; 1; 1; \ldots; 1; 2-1/n]$, жорданова нормальная форма матрицы W диагональная, собственный вектор, отвечающий собственному значению 0, равен $\mathrm{v}_0 = (1, 1, \ldots, 1)^T$, а максимальному собственному значению —

$$\mathrm{v}_{\max} = \mathrm{v}_0 - \left(2 - \frac{1}{n}\right) \underbrace{(1, 0, 0, \ldots, 0)^T}_{e_1}$$

и собственное подпространство, отвечающее собственному значению 1, ортогонально v_0 и e_1, а потому $\widetilde{\sigma}_{\min}(W) = 1$ и

$$\sigma_{\max}(W) = \max_{\|h\|_2=1} \|Wh\|_2^2 = \|We_1\|_2^2 \approx n$$

(см. [95, 304, 585]). Таким образом, отказавшись от симметричности матрицы, в рассматриваемом случае можно уменьшить $\sigma_{\max}(W)/\widetilde{\sigma}_{\min}(W)$ в $\sim n$ раз, однако это не позволяет ускорить сходимость градиентного спуска, поскольку в случае несимметричной матрицы W уже нельзя работать, как описано выше, с $\sqrt{W}$.

Заметим, что в общем случае задача наилучшего подбора системы линейных уравнений, равносильной системе $y_1 = \ldots = y_n$ (и соответствующей этой системе матрицы W), тесно связана с задачами выпуклого полуопределённого программирования, рассмотренными в работе [546]. Другой способ подбора см. в комментарии к упражнению 5.9 (конструкция предобуславливателя).

Отметим также популярную в последнее время тенденцию (см., например, [247, 248, 441, 541, 579] и замечание 1.5): сведение задачи поиска наилучшего численного метода для рассматриваемого класса задач оптимизации, в свою очередь, к задаче оптимизации (седловой задаче). Если удаётся решить такую (мета-)задачу, то удаётся найти наилучший метод (в минимаксном смысле). Однако, как правило, точное и удобное для практического применения решение удаётся найти в редких случаях (но всё же удаётся, см. замечание 1.5), поэтому интересны и различные релаксации (упрощения) сложной метазадачи. ♦

Упражнение 4.8 (децентрализованная распределённая оптимизация в сильно выпуклом случае [307, 540]). Покажите, что если в примере 4.1 дополнительно предположить, что $\varphi_i''(y) \leqslant L_\varphi$, то «двойственная» функция $f(x)$, определённая соотношением (4.20), будет сильно выпуклой в 2-норме с константой $\mu_f = \widetilde{\sigma}_{\min}(W)/L_\varphi$ на $(\operatorname{Ker} W^T)^\perp = (\operatorname{Ker} W)^\perp$ (при редуцированном подходе $\mu_f = \widetilde{\sigma}_{\min}(\sqrt{W})/L_\varphi$ и на $(\operatorname{Ker}\sqrt{W})^\perp$).

Указание. См. [255, утверждение 2.1], [485, теорема 1], [373, теорема 6], [437, лемма 3.1], [458] , [527, теорема 23.5]. Следует обратить внимание на неполную симметричность следующих связей: 1) сильная выпуклость прямой задачи порождает гладкость (липшицевость градиента) двойственной и 2) гладкость в прямой задаче порождает сильную выпуклость двойственной. Во втором случае требуется, чтобы при переходе к двойственной задаче все ограничения с помощью множителей Лагранжа переносились в функционал. В первом случае этого не требуется. Эта несимметричность вместе с теоремой Фенхеля — Моро [66, п. 1.4, 2.2] отчасти объясняет отмеченную ранее асимметрию в возможности адаптивной настройки методов на неизвестную константу Липшица градиента и отсутствие такой возможности для константы сильной выпуклости. Во всяком случае, пока не придумали, как можно было бы в общем случае адаптивно осуществлять такую настройку без серьёзных дополнительных усилий, см. указание к упражнению 1.3.

♦ Отметим, что сильная выпуклость двойственного функционала $f(x) = \varphi^*(-Wx)$ (см. формулу (4.20)) имеет место только на $(\operatorname{Ker} W)^\perp$, см. упражнение 4.8. Точнее говоря, сильная выпуклость имеет место на любой гиперплоскости вида $x^0 + (\operatorname{Ker} W)^\perp$. При этом важно заметить, что

$$\nabla f(x) = -W^T \nabla \varphi^*(z)\big|_{z=-Wx} \in \mathfrak{J}(-W^T) = \mathfrak{J}W^T = (\operatorname{Ker} W)^\perp,$$

т. е. $x^k \in x^0 + (\operatorname{Ker} W)^\perp$ для любого k. Таким образом, траектория градиентного спуска всё время будет находиться в той же самой гиперплоскости $x^0 + (\operatorname{Ker} W)^\perp$, в которой функция $f(x)$ является μ_f-сильно выпуклой. Последнее означает, что для градиентного спуска (аналогично и для быстрого градиентного спуска) двойственную задачу можно считать μ_f-сильно выпуклой, «забывая» про вырождение на подпространстве $\operatorname{Ker} W$.

Изменяющиеся со временем коммуникационные графы стали достаточно популярными в последнее время в связи с ростом интереса к изучению различных беспроводных систем мобильных объектов с ограниченным радиусом действия антенн. ♦

Замечание 4.4 (распределённая оптимизация и модели консенсуса). Заметим, что распределённое решение задач оптимизации можно осуществлять, не переходя к двойственной задаче. Поясним идею прямого подхода. Пусть матрица Лапласа неориентированного связного графа рассматриваемой коммуникационной сети равна W, см. пример 4.1. В i-м узле графа хранится число x_i^0, $i = 1, \ldots, n$. Требуется так организовать коммуникацию[8], чтобы система сошлась к консенсусу [1, 625] за наименьшее число коммуникационных шагов. Поставим в соответствие этой задаче следующую задачу квадратичной оптимизации:

$$\frac{1}{2}\langle x, Wx\rangle \to \min_{x \in \mathbb{R}^n}.$$

Очевидно, что решением этой задачи будет любой вектор вида $x_* = \operatorname{const} \cdot (1, \ldots, 1)^T$. Рассмотрим быстрый градиентный метод из указания к упражнению 1.3, который, подобно примеру 4.1, имеет естественную интерпретацию с точки зрения коммуникации узлов с прямыми соседями[9]:

$$x^1 = x^0 - \frac{1}{\lambda_{\max}(W)} W x^0,$$

$$x^{k+1} = x^k - \frac{1}{\lambda_{\max}(W)} W \cdot \left(x^k + \frac{\sqrt{\lambda_{\max}(W)} - \sqrt{\tilde{\lambda}_{\min}(W)}}{\sqrt{\lambda_{\max}(W)} + \sqrt{\tilde{\lambda}_{\min}(W)}} (x^k - x^{k-1}) \right) + \\ + \frac{\sqrt{\lambda_{\max}(W)} - \sqrt{\tilde{\lambda}_{\min}(W)}}{\sqrt{\lambda_{\max}(W)} + \sqrt{\tilde{\lambda}_{\min}(W)}} (x^k - x^{k-1}),$$

[8] По условию на одном шаге коммуникации каждый узел может обмениваться информацией только со своими прямыми соседями.

[9] Заметим, что стандартный метод градиентного спуска (см. § 1) для рассматриваемой задачи можно понимать как метод простой итерации [95, п. 3.4]. Этот метод будет иметь ещё более простую интерпретацию.

где $\tilde{\lambda}_{\min}(W) = \widetilde{\sigma}_{\min}(\sqrt{W})$ — наименьшее из положительных собственных значений матрицы W. Аналогично комментарию к упражнению 4.8 можно показать, что этот метод сходится к такому x_*, что $x_* \in x^0 + (\operatorname{Ker} W)^{\perp}$, т. е. к *консенсусу* (состоянию, когда все узлы «знают» среднее арифметическое начальных состояний узлов):

$$x_* = \frac{1}{n}\sum_{i=1}^{n} x_i^0 \cdot (1, \ldots, 1)^T.$$

При этом для того, чтобы выполнялась оценка

$$\|x^N - x_*\|_2 \leqslant \varepsilon \|x^0 - x_*\|_2,$$

достаточно $N = \mathrm{O}\left(\sqrt{\chi(W)} \ln \varepsilon^{-1}\right)$ итераций (коммуникационных шагов)[10], где

$$\chi(W) = \frac{\lambda_{\max}(W)}{\tilde{\lambda}_{\min}(W)}.$$

Если теперь считать, что вместо x_i^0, $i = 1, \ldots, n$, в узлах хранятся (вычисляются) $\nabla\varphi_i(y)$, $i = 1, \ldots, n$, то аналогичным образом за дополнительную мультипликативную плату $\mathrm{O}\left(\sqrt{\chi(W)} \ln \varepsilon^{-1}\right)$ можно добиться того, чтобы все узлы «знали» градиент всего функционала

$$\varphi(y) = \sum_{i=1}^{n} \varphi_i(y) \to \min_y.$$

Таким образом[11], если каждая функция $\varphi_i(y)$ в исходной задаче оптимизации имеет L_φ-липшицев градиент и константу сильной выпуклости μ_φ, а вместо базового градиентного метода используется (и в прямом, и в двойственном подходе) быстрый градиентный метод, то общее число коммуникационных шагов при прямом подходе будет равно

$$\mathrm{O}\left(\sqrt{\frac{nL_\varphi}{n\mu_\varphi}} \ln \varepsilon^{-1} \cdot \sqrt{\chi(W)} \ln \varepsilon^{-1}\right) = \mathrm{O}\left(\sqrt{\frac{L_\varphi}{\mu_\varphi}} \cdot \sqrt{\chi(W)} \ln^2 \varepsilon^{-1}\right),$$

[10] Считаем, что $\chi(W) \ll \varepsilon^{-1}$.

[11] На самом деле эти рассуждения не точны и приведены здесь лишь для наглядности восприятия материала. В действительности здесь требуется дополнительно пенализировать исходную постановку задачи квадратичным штрафом, описанным в конце замечания 4.2, но общая связь с консенсусным алгоритмом сохраняется [262, 307, 428, 430]. Впрочем, имеются и прямые рассуждения [530, 614], максимально приближенные к описанной процедуре. Отметим, что проверить корректность доказательств из работы [614] пока не получилось.

где ε — желаемая относительная точность решения задачи по аргументу, а при двойственном подходе —

$$O\left(\sqrt{\frac{L_\varphi}{\mu_\varphi}\cdot\chi(W)}\ln\varepsilon^{-1}\right),$$

см. упражнения 1.3, 4.8. При прямом подходе число вызовов градиентного оракула в каждом узле будет равно

$$O\left(\sqrt{\frac{nL_\varphi}{n\mu_\varphi}}\ln\varepsilon^{-1}\right)=O\left(\sqrt{\frac{L_\varphi}{\mu_\varphi}}\ln\varepsilon^{-1}\right),$$

а для двойственного подхода (см. формулу (4.20) с $W\to\sqrt{W}$, а также дальнейшие рассуждения в этом замечании) число вызовов двойственного градиентного оракула будет составлять

$$O\left(\sqrt{\frac{L_\varphi}{\mu_\varphi}\cdot\chi(W)}\ln\varepsilon^{-1}\right).$$

Стоит отметить, что в прямом подходе возможно дополнительное ускорение за счёт следующей простой конструкции [170, 540]. Строим остовное дерево для исходной коммуникационной сети. В принципе, это можно сделать и распределённым образом [289, 540]. Объявляем одну из вершин корнем. Выделение корня порождает иерархию родитель → потомок. Цель та же самая, что и раньше. На первом шаге коммуникации каждый узел, у которого нет потомков (лист дерева), посылает своё значение (градиент) своему родителю. Родители прибавляют полученные от потомков значения к своему. Система переходит на следующий шаг коммуникации, в котором роль листьев будут играть родители, получившие на прошлом шаге коммуникации значения от своих предков. Процесс повторяется до тех пор, пока вся сумма значений (во всех узлах) не соберётся в корневом узле. Корневой узел, базируясь на градиенте всего функционала, может сделать шаг выбранного итерационного метода и полученное(-ые) в результате значение(-я) распространить среди потомков, запустив процесс в обратном порядке. При самом неблагоприятном выборе корня дерева число коммуникационных шагов при таком подходе будет не больше удвоенного диаметра исходного коммуникационного графа. При этом во всех случаях этот диаметр будет не больше, чем корень из числа обусловленности матрицы Лапласа этого графа $\sqrt{\chi(W)}$. Часто это величины одного порядка [540]. Таким образом, появляется возможность «отыграть» потерю в логарифмическом множителе в оценке

числа коммуникационных шагов при прямом подходе. Более того, для рассмотренной в комментарии к упражнению 4.7 звёздной топологии диаметр графа оказывается в $\sim\sqrt{n}$ меньше корня из числа обусловленности. Таким образом, дополнительный выигрыш от второго подхода может быть и существенным[12].

Отсюда, однако, не стоит делать вывод о том, что прямой подход всегда лучше двойственного. Во-первых, стоимость прямого и двойственного градиентного оракула может быть существенно разной [264, 591], и выбор в пользу одного из подходов следует осуществлять, учитывая и стоимость вызова соответствующего оракула. Действительно, в задаче вычисления барицентра Монжа — Канторовича — Васерштейна (МКВ) [508] n вероятностных дискретных мер (с носителем на d точках) вместо настоящего расстояния МКВ считают μ-энтропийно регуляризованное расстояние, где $\mu = \varepsilon/(2n\ln d^2)$, см. замечание 4.1, поэтому каждое слагаемое в сумме (4.17) будет иметь вид

$$\varphi_l(y) = \varphi_{w^l}(y) = \min_{\substack{\sum_{j=1}^d \pi_{ij}=y_i,\ \sum_{i=1}^d \pi_{ij}=w_j^l \\ \pi_{ij}\geqslant 0,\ i,j=1,\ldots,d}} \left\{ \sum_{i,j=1}^d c_{ij}\pi_{ij} + \mu \sum_{i,j=1}^d \pi_{ij}\ln\pi_{ij} \right\} =$$

$$= \max_{x\in\mathbb{R}^d} \min_{\substack{\sum_{i=1}^d \pi_{ij}=w_j^l \\ \pi_{ij}\geqslant 0,\ i,j=1,\ldots,d}} \left\{ \sum_{i,j=1}^d c_{ij}\pi_{ij} + \sum_{i=1}^d x_i\cdot\left(y_i - \sum_{j=1}^d \pi_{ij}\right) + \mu\sum_{i,j=1}^d \pi_{ij}\ln\pi_{ij} \right\} =$$

$$= \max_{x\in\mathbb{R}^d} \Bigg\{ \langle y, x\rangle \underbrace{- \mu \sum_{j=1}^d w_j^l \ln\left(\frac{1}{w_j^l} \sum_{i=1}^d \exp\left(\frac{-c_{ij}+x_i}{\mu}\right)\right)}_{\varphi_l^*(x)=\varphi_{w^l}^*(x)} \Bigg\},$$

где $l=1,\ldots,n$, $y\in S_d(1)$. Достаточно точный подсчёт $\varphi_{w^l}(y)$ и $\nabla\varphi_{w^l}(y)$ при малых значениях ε требует $\widetilde{O}(d^3)$ арифметических операций [131, 287, 506, 577] (см. также указание к упражнению 1.4 и указание к упражнению 3.9, в котором отмечается, что при не малых значениях ε существуют более эффективные способы), в то время как подсчёт $\varphi_{w^l}^*(y)$ и $\nabla\varphi_{w^l}^*(y)$ — всего лишь $O(d^2)$. Отметим также, что двойственная задача — задача безусловной оптимизации, а прямая задача — условной, поскольку есть ограничение $y\in S_d(1)$. Однако наличие ограничения простой структуры не скажется на возможности применения прямого подхода.

[12] К сожалению, второй подход, требующий вычисления остовного дерева, плохо подходит для работы на изменяющихся со временем графах.

Во-вторых, для негладких, но сильно выпуклых прямых задач[13] прямой подход требует

$$\widetilde{\mathrm{O}}\Big(\frac{M^2}{n\mu\cdot\varepsilon}\cdot\sqrt{\chi(W)}\Big)$$

коммуникационных шагов и

$$\mathrm{O}\Big(\frac{M^2}{n\mu\cdot\varepsilon}\Big)$$

вызовов прямого градиентного оракула в каждом узле (см. упражнение 2.3 и табл. 2 в приложении), где M — константа Липшица прямого функционала $\varphi(y)=\sum_{l=1}^n\varphi_{w^l}(y)$, $y\in\widetilde{Q}\subseteq\mathbb{R}^d$. В двойственном подходе к задаче

$$\widetilde{\varphi}(Y=(y_1,\ldots,y_n))=\sum_{l=1}^n\varphi_{w^l}(y_l)\to\min_{\substack{\sqrt{W}Y=0\\ y_i\in\widetilde{Q},\,l=1,\ldots,n}}$$

строится (с точностью до знака) двойственный функционал, который необходимо минимизировать (см. (4.20) с $W\to\sqrt{W}$):

$$f(x)=\sum_{l=1}^n\varphi_l^*([-\sqrt{W}x]_l)\to\min_{x\in\mathbb{R}^n},$$

где в выписанных формулах $\sqrt{W}:=\sqrt{W}\otimes 1_d1_d^T$, а $\otimes$ — кронекерово произведение матриц [405]. Число коммуникационных шагов и число вызовов оракула, выдающего градиент соответствующей (рассматриваемому узлу) двойственной функции, совпадут (это общее положение для двойственного оракула, см., например, полученные выше оценки в гладком сильно выпуклом случае) и будут равны

$$\mathrm{O}\left(\sqrt{\frac{L_fR_x^2}{\varepsilon}}\right)=\mathrm{O}\left(\sqrt{\frac{1}{\varepsilon}\left(\frac{\sigma_{\max}(\sqrt{W})}{\mu}\right)\left(\frac{\|\nabla\widetilde{\varphi}(Y_*)\|_2^2}{\widetilde{\sigma}_{\min}(\sqrt{W})}\right)}\right)=$$

$$=\mathrm{O}\left(\sqrt{\frac{\|\nabla\widetilde{\varphi}(Y_*)\|_2^2}{\mu\varepsilon}\cdot\chi(W)}\right),$$

см. указание к упражнению 1.3, замечание 4.2, пример 4.1 и табл. 2 в приложении. Для того чтобы сравнить приведённые оценки, заметим, что

$$M^2=\max_{y\in\widetilde{Q}\cap B_{2R}(y^0)}\|\nabla\varphi(y)\|_2^2\quad\text{и}\quad\|\nabla\varphi(y_*)\|_2^2\leqslant n\|\nabla\widetilde{\varphi}(Y_*)\|_2^2.$$

[13] Такой задачей как раз будет задача вычисления μ-энтропийно регуляризованного барицентра МКВ — $\varphi_l(y)$ будет μ-сильно выпуклой функцией в 2-норме.

Из этих оценок в общем случае можно сделать довольно грубый вывод о том, что

$$\frac{M^2}{n} \approx \|\nabla\widetilde{\varphi}(Y_*)\|_2^2.$$

В таком приближении при двойственном подходе оценки на число коммуникаций и вызовов оракула (в каждом узле) получаются извлечением корня из оценок в прямом подходе.

Отметим при этом, что на основе двойственного подхода можно предложить такой способ решения исходной задачи распределённым образом с прямым оракулом, т. е. оракулом, выдающим градиент прямых функций (в соответствующих узлах), при котором число коммуникационных шагов будет такое же, как в двойственном подходе. Для этого необходимо использовать двойственный подход, а для вычисления градиента двойственной функции (соответствующей рассматриваемому узлу) использовать прямой оракул. А именно, чтобы с нужной точностью вычислить

$$\nabla\varphi_l^*([-\sqrt{W}x]_l) = -\sqrt{W}y_l(x),$$

нужно с помощью прямого оракула достаточно точно решить задачу

$$y_l(x) = \arg\max_{y_l\in Q}\left\{\langle[-\sqrt{W}x]_l, y_l\rangle - \varphi_l(y_l)\right\}.$$

К сожалению, при таком подходе без дополнительных предположений[14] не удаётся получить ожидаемые (из прямого подхода) оценки на число вызовов прямого оракула[15]. Оказывается, можно предложить такой распределённый способ решения исходной задачи [307, 405], который обеспечивает оптимальное число коммуникаций и одновременно ожидаемое (из прямого подхода) число вызовов оракула в общем случае, причём как в случае выпуклого функционала, так и в рассмотренном выше случае сильно выпуклого (прямого) функционала. Для этого стоит обратиться к конструкции, изложенной в комментарии к замечанию 4.3 (стоит иметь в виду, что обозначения силь-

[14] Например, таких, что для регуляризованной двойственной функции доступен градиент сопряжённой (двойственной) к ней функции, см. замечание 4.1 в части двойственного сглаживания. Таким образом, доступен не прямой оракул, а двойственно-сглаженный прямой оракул.

[15] Здесь говорится об оптимальном числе вызовов прямого оракула (на узле) и оптимальном числе коммуникаций [142, 262, 340, 539, 540]. Отметим работы [261, 339–342, 390, 414, 429, 605, 609], в которых обсуждаются вопросы возможности распространения ускорения из замечания 1 приложения на распределённые (параллельные) алгоритмы, в предположении, что на каждом узле хранится часть (больше чем одно слагаемое) исходной суммы.

но отличаются). В этой конструкции в выпуклом случае описывается способ получения оптимальных оценок на число коммуникационных шагов и «ожидаемых» оценок на число вызовов прямого оракула. С помощью рестартов (см. указание к упражнении 2.3 и конец § 5) можно перенести эти результаты и на сильно выпуклый случай, получив анонсированный результат.

Заметим, что оптимальные оценки на число коммуникационных шагов в случае не сильно выпуклого функционала прямой задачи можно получать из разобранных выше двух случаев просто за счёт регуляризации исходной (прямой) задачи, см. замечание 4.1. При этом двойственный оракул должен будет выдавать градиент не сопряжённой функции к прямой, а сопряжённой функции к регуляризованной прямой. Отметим также, что в гладком случае «ожидаемые» оценки (с точностью до логарифмического множителя) на число вызовов прямого оракула получаются после регуляризации из разобранного в самом начале случая (гладкого и сильно выпуклого), см. также [307, 428]. ■

♦ Отметим, что приведённые выше результаты частично переносятся и на меняющиеся со временем графы [530, 531, 574]. Можно отказаться и от неориентированности коммуникационного графа / симметричности матрицы W, сохраняя при этом его связность. В этом случае также можно получить результаты, частично похожие на те, что были приведены в пособии, см., например, работы [459, 574] и цитированную там литературу. К сожалению, даже если рассматривать только ориентированные, не изменяющиеся со временем коммуникационные графы, то на данный момент неизвестно, можно ли (а если можно, то каким образом) ускорить сходимость, как это было сделано в случае неориентированных графов [307, 540]. Впрочем, существует достаточно общий способ ускорения, в том числе и распределённых централизованных алгоритмов, — каталист (см. замечание 3.3 и приложение), в котором предлагается решать (внутреннюю) вспомогательную (сильно выпуклую) задачу рассматриваемым неускоренным распределённым централизованным методом. В частности, таким образом можно, например, ускорить асинхронный централизованный алгоритм из работы [448]. ♦

Упражнение 4.9 (схема регуляризации А. Н. Тихонова [8, лекции 5, 7]). Пусть необходимо решить систему линейных уравнений $Ax = b$ *в условиях истокопредставимости*:

$$x_* = A^T y_*, \quad \|y_*\|_2 \leqslant R.$$

Как будет видно из указания, условие истокопредставимости можно также понимать следующим образом (см. замечание 1.6):

$$x_* = (A^TA)^{1/2}\tilde{y}_*, \quad \|\tilde{y}_*\|_2 \leqslant R.$$

Результат не изменится.

Предположим, что точное значение правой части b неизвестно, зато известно такое значение $\tilde{b}$, что $\|\tilde{b}-b\|_2 \leqslant \delta$. Схема А. Н. Тихонова предлагает искать решение системы $Ax = b$ путём решения задачи оптимизации

$$\frac{1}{2}\|Ax-\tilde{b}\|_2^2 + \frac{\mu}{2}\|x\|_2^2 \to \min_{x\in\mathbb{R}^n}.$$

Считая, что известно точное решение этой задачи

$$x_*^\mu(\tilde{b}) = (A^TA+\mu I)^{-1}A^T\tilde{b},$$

предложите способ выбора параметра $\mu(\delta)$ так, чтобы в оценке

$$\|x_* - x_*^\mu(\tilde{b})\|_2 \leqslant \sigma(\delta)$$

функция $\sigma(\delta) \geqslant 0$ была как можно меньше.

Указание. По неравенству треугольника

$$\|x_* - x_*^\mu(\tilde{b})\|_2 \leqslant \|x_* - x_*^\mu(b)\|_2 + \|x_*^\mu(b) - x_*^\mu(\tilde{b})\|_2.$$

Заметим, что

$$x_* - x_*^\mu(b) = (A^TA+\mu I)^{-1}(A^TA+\mu I)x_* - (A^TA+\mu I)^{-1}\underbrace{A^Tb}_{A^TAx_*} =$$

$$= \mu(A^TA+\mu I)^{-1}x_* = \mu\underbrace{(A^TA+\mu I)^{-1}A^T}_{\Theta(A,\mu)}y_*,$$

$$x_*^\mu(b) - x_*^\mu(\tilde{b}) = (A^TA+\mu I)^{-1}A^T(b-\tilde{b}) = \Theta(A,\mu)(b-\tilde{b}).$$

Таким образом, необходимо оценить норму оператора $\Theta(A,\mu)$:

$$\|\Theta(A,\mu)\|_2 = \sqrt{\sup_{\|z\|_2=1}\|\Theta(A,\mu)z\|_2^2}.$$

Пусть $\{e_k\}_k$ — ортонормированный базис пространства из собственных векторов самосопряжённого (симметричного) оператора (матрицы) A^TA. Тогда по спектральной теореме [61]

$$\|\Theta(A,\mu)z\|_2^2 = \sum_k \frac{1}{(\lambda_k+\mu)^2}\langle e_k, A^Tz\rangle^2 = \sum_{k:\,\lambda_k>0}\frac{\lambda_k}{(\lambda_k+\mu)^2}\left\langle \frac{1}{\sqrt{\lambda_k}}Ae_k, z\right\rangle^2.$$

Поскольку векторы

$$\left\{\frac{1}{\sqrt{\lambda_k}}Ae_k\right\}_{k:\,\lambda_k>0}$$

образуют ортонормированную систему (но не обязательно базис), по неравенству Бесселя имеем

$$\sum_k\left\langle\frac{1}{\sqrt{\lambda_k}}Ae_k, z\right\rangle^2 \leqslant \|z\|_2^2.$$

Следовательно,

$$\|\Theta(A,\mu)z\|_2^2 \leqslant \sup_k \frac{\lambda_k}{(\lambda_k+\mu)^2}\|z\|_2^2.$$

Значит,

$$\|\Theta(A,\mu)\|_2 \leqslant \sup_k \frac{\sqrt{\lambda_k}}{\lambda_k+\mu} \leqslant \sup_{\lambda\geqslant 0}\frac{\sqrt{\lambda}}{\lambda+\mu} = \frac{1}{2\sqrt{\mu}}.$$

Таким образом,

$$\|x_* - x_*^\mu(\tilde{b})\|_2 \leqslant \|x_* - x_*^\mu(b)\|_2 + \|x_*^\mu(b) - x_*^\mu(\tilde{b})\|_2 \leqslant \frac{R\sqrt{\mu}}{2} + \frac{\delta}{2\sqrt{\mu}}.$$

Правая часть последнего неравенства минимальна при выборе $\mu(\delta) = \delta/R$. В таком случае

$$\|x_* - x_*^\mu(\tilde{b})\|_2 \leqslant \sqrt{R\delta} = \sigma(\delta).$$

К сожалению, на практике R обычно неизвестно. Есть сложности и с предположением об известности δ (вето Бакушинского [332]). Альтернативным и часто более эффективным способом решения такого рода задач является регуляризация за счёт правильного выбора численного метода решения нерегуляризованной задачи

$$\frac{1}{2}\|\tilde{A}x - \tilde{b}\|_2^2 \to \min_{x\in\mathbb{R}^n},$$

обладающего регуляризирующими свойствами, см. замечание 1.6.

§ 5

Универсальный градиентный спуск

Как и в § 2–4, рассмотрим общую задачу выпуклой оптимизации (2.1): $f(x) \to \min_{x \in Q}$.

В данном параграфе, следуя Ю. Е. Нестерову [492] (см. также [7, 31, 32, 97, 296, 470, 615]), мы сделаем, пожалуй, самый важный шаг во всём описанном выше подходе — согласуем формулы (2.5), (2.26), (2.27). Как уже отмечалось ранее, для наглядности рассуждения будут проводиться не в максимальной общности (см. § 3).

Прежде всего заметим, что во всех вариантах рассмотренных на данный момент методов градиентного спуска мы использовали шаг $h = 1/L$, где константа L либо была задана по условию, либо определялась согласно соотношению (2.5) с $\delta = \varepsilon/2$ (см. вывод формул (4.6), (4.28)). Рассмотрим следующее (универсальное) обобщение метода (2.28) (описывается k-я итерация).

Универсальный градиентный спуск

$L^{k+1} = \dfrac{L^k}{2}$

While True Do

$x^{k+1} = \arg\min_{x \in Q} \left\{ f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + L^{k+1} V(x, x^k) \right\}$

If $\left\{ f(x^{k+1}) \leqslant f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + L^{k+1} V(x^{k+1}, x^k) + \dfrac{\varepsilon}{2} \right\}$

Перейти на следующую итерацию: $k \to k+1$

Else

$L^{k+1} := 2L^{k+1}$

Для такого метода формула (4.1) перепишется следующим образом:

$$\frac{1}{L^{k+1}} f(x^{k+1}) \leqslant \frac{1}{L^{k+1}} \left\{ f(x^k) + \langle \nabla f(x^k), x - x^k \rangle \right\} + \\ + V(x, x^k) - V(x, x^{k+1}) + \frac{\varepsilon}{2L^{k+1}}, \quad (5.1)$$

где константа L^{k+1} подбирается согласно описанной выше процедуре. При этом согласно неравенству (2.32) эта константа оценивается сверху константой L^{k+1}, подбираемой из соотношения (2.26), в котором $\delta = \varepsilon/2$:

$$f(x^{k+1}) \leqslant f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L^{k+1}}{2}\|x^{k+1} - x^k\|^2 + \frac{\varepsilon}{2}.$$

Таким образом, автоматически происходит подбор на рассматриваемом отрезке $[x^k, x^{k+1}]$ двух параметров ν и L_ν в неравенство (2.27):

$$\|\nabla f(y) - \nabla f(x)\|_* \leqslant L_\nu \|y - x\|^\nu, \quad \nu \in [0,1],\ L_0 < \infty,$$

так, чтобы неравенство (2.26) выполнилось. Из соотношения (2.5) и описанной выше процедуры подбора L^{k+1} следует, что

$$L^{k+1} \leqslant L_\nu \cdot \Big[\frac{L_\nu}{\varepsilon}\frac{1-\nu}{1+\nu}\Big]^{(1-\nu)/(1+\nu)}.$$

Подчеркнём, что не мы сами решаем задачу подбора ν и L_ν — это делает метод за счёт описанной процедуры. Свойства гладкости функции $f(x)$ на отрезке $[x^k, x^{k+1}]$ характеризуются континуальным набором чисел $\{L_\nu\}_{\nu\in[0,1]}$, часть из которых может равняться бесконечности. Мы можем ничего не знать о $\{L_\nu\}_{\nu\in[0,1]}$ — при универсальном подходе этого и не требуется. Тем не менее описанная выше процедура гарантирует, что метод подберёт такое $\nu \in [0,1]$, что соответствующая этому ν константа Гёльдера L_ν порождает (по формуле (2.26) с $\delta = \varepsilon/2$) на отрезке $[x^k, x^{k+1}]$ минимально возможную (с точностью до множителя, не большего 2) константу L^k, которая явно используется в методе. Подобно оценкам (4.6), (4.28), для универсального градиентного спуска можно получить следующий результат.

Теорема 5.1. *Пусть нужно решить задачу* (2.1) *в условиях* (2.27). *Для универсального градиентного спуска после*[1]

$$N = \inf_{\nu\in[0,1]} \Big(\frac{2L_\nu R^{1+\nu}}{\varepsilon}\Big)^{2/(1+\nu)} \tag{5.2}$$

[1] При этом в формуле (5.2) выбираются константы L_ν, не худшие, по всему пространству, а «средние» на траектории метода. При получении оценки (5.2) мы предполагали, что L^0 выбрано так, что при подстановке $k=0$ в условие «**If**» рассматриваемого универсального метода неравенство в этом условии будет неверным при $L^{k+1} = L^1 < L^0$. На самом деле, если предположение о выборе L^0 не выполнено, это приведёт лишь к небольшому ухудшению числового множителя в оценке (5.2).

итераций имеет место следующая оценка:

$$f(\bar{x}^N) - f(x_*) \leqslant f(\bar{x}^N) - \\ - \frac{1}{\sum_{k=0}^{N-1} \frac{1}{L^{k+1}}} \min_{x \in B_{R,Q}(x^0)} \left\{ \sum_{k=0}^{N-1} \frac{1}{L^{k+1}} [f(x^k) + \langle \nabla f(x^k), x - x^k \rangle] \right\} \leqslant \varepsilon,$$

где (*в данном случае*)

$$\bar{x}^N = \frac{1}{\sum_{k=1}^{N} \frac{1}{L^k}} \sum_{k=1}^{N} \frac{x^k}{L^k},$$

$R^2 = V(x_*, x^0)$. *Если решение* x_* *не единственно, то оценка* (5.2) *справедлива и для того решения* x_*, *которое доставляет минимум* R^2.

Ещё раз подчеркнём, что задачу минимизации, возникающую в оценке (5.2), нам решать не нужно, равно как и не нужно знать хоть что-то о гладкости функции $f(x)$, кроме того, что $L_0 < \infty$, чтобы универсальный метод сходился. В негладком случае (когда только $L_0 < \infty$) имеем

$$N = \frac{4L_0^2 R^2}{\varepsilon^2},$$

что с точностью до множителя соответствует нижней оценке, см. формулу (2.37).

Рассуждая аналогично [492], несложно показать, что описанный выше универсальный метод на каждой итерации запрашивает один раз $\nabla f(x^k)$ и в среднем (по итерациям) около трёх раз значение функции $f(x)$. Действительно, среднее число вычислений значения функции $f(x)$ есть

$$\frac{1}{N} \sum_{k=0}^{N-1} \left(2 + \log_2 \left(\frac{L^{k+1}}{L^k/2}\right)\right) = 3 + \frac{1}{N} \log_2 \left(\frac{L^N}{L^0}\right).$$

Замечание 5.1 (универсальный метод для вариационных неравенств и седловых задач[2] **[22, 265])**. Аналогом градиентного метода

[2] В концепции модели функции из примера 3.1 на базе этого замечания можно получить универсальный вариант популярного сейчас *метода Шамболя — Пока* [203, 444, 568]. Отметим, что в работах [444, 445] предложен альтернативный способ (forward–backward–forward) построения адаптивных методов решения вариационных неравенств. Этот способ использует только одну проекцию на каждой итерации и при этом имеет аналогичные оценки скорости сходимости. Похожий на описанный здесь алгоритм для более общего класса задач (монотонных включений) описан в работе [238].

для вариационных неравенств (ВН) и седловых задач является *экстраградиентный метод* Г. М. Корпелевич [13, § 15, гл. 5], [64, 148]. Интересные результаты по решению ВН можно найти в работах А. С. Антипина [136], см. также обзор в работе [149]. Далее рассмотрим один современный вариант экстраградиентного метода, а именно *проксимальный зеркальный метод* А. С. Немировского [466]. Пусть задано векторное поле $g(x)$, в частности $g(x) = \nabla f(x)$. Предположим, что существуют такие L и δ, что для всех x, y, z из выпуклого компактного множества Q имеет место неравенство [216, 569]

$$\langle g(y) - g(x), y - z\rangle \leqslant LV(y,x) + LV(y,z) + \delta.$$

Тогда для проксимального зеркального метода

$$y^{k+1} = \arg\min_{x\in Q}\left\{\langle g(x^k), x - x^k\rangle + LV(x,x^k)\right\},$$
$$x^{k+1} = \arg\min_{x\in Q}\left\{\langle g(y^{k+1}), x - x^k\rangle + LV(x,x^k)\right\}$$

имеет место следующая оценка:

$$\frac{1}{N}\sum_{k=1}^{N}\langle g(y^k), y^k - x\rangle \leqslant \frac{LV(x,x^0) - LV(x,x^N)}{N} + \delta. \tag{5.3}$$

С помощью текста, написанного в конце п. 4.6 работы [186], несложно построить универсальный вариант такого метода (описывается k-я итерация).

Универсальный проксимальный зеркальный метод

$L^{k+1} = \dfrac{L^k}{2}$

While True Do

$y^{k+1} = \arg\min_{x\in Q}\left\{\langle g(x^k), x - x^k\rangle + L^{k+1}V(x,x^k)\right\},$

$x^{k+1} = \arg\min_{x\in Q}\left\{\langle g(y^{k+1}), x - x^k\rangle + L^{k+1}V(x,x^k)\right\}$

If $\left\{\langle g(y^{k+1}) - g(x^k), y^{k+1} - x^{k+1}\rangle \leqslant L^{k+1}V(y^{k+1},x^k) + L^{k+1}V(y^{k+1},x^{k+1}) + \dfrac{\varepsilon}{2}\right\}$

Перейти на следующую итерацию: $k \to k+1$

Else

$L^{k+1} := 2L^{k+1}$

Если, подобно неравенству (2.27), векторное поле $g(x)$ удовлетворяет условию

$$\|g(y) - g(x)\|_* \leqslant L_\nu \|y - x\|^\nu, \quad \nu \in [0,1],\ x, y \in Q,\ L_0 < \infty,$$

то, используя неравенство (верное для любых $a, b, L_\nu, \delta > 0$, $\nu \in [0,1]$)

$$L_\nu a^\nu b \leqslant L_\nu \cdot \Big(\frac{L_\nu}{\delta}\Big)^{(1-\nu)/(1+\nu)} \Big(\frac{a^2}{2} + \frac{b^2}{2}\Big) + \delta$$

с $\delta = \varepsilon/2$, можно получить (см. [22]), что для достижения оценки

$$\frac{1}{\sum_{k=1}^N \frac{1}{L^k}} \max_{x\in Q} \Big\{ \sum_{k=1}^N \frac{1}{L^k} \langle g(y^k), y^k - x\rangle \Big\} \leqslant \varepsilon \tag{5.4}$$

достаточно (следует сравнить с соотношением (5.2))

$$N = \inf_{\nu\in[0,1]} \Big(\frac{2L_\nu R^{1+\nu}}{\varepsilon}\Big)^{2/(1+\nu)} \tag{5.5}$$

итераций универсального проксимального зеркального метода. Здесь $R^2 = \max_{x\in Q} V(x, x^0)$. При этом среднее число вычислений значений векторного поля $g(x)$ на одной итерации приближённо равно трём. Оценка (5.5) с точностью до числового множителя оптимальна для ВН и для седловых задач при[3] $N \leqslant \dim x$. К сожалению, точной ссылки на обоснование оптимальности найти не удалось, однако различные частные случаи могут быть сведены к разобранному в работах [74, 326, 462, 463, 502].

♦ Следует сравнить оценку (5.4) при $g(x) = \nabla f(x)$ с (2.25), (3.4). Если

$$g(x) = (\nabla_u f(u,w), -\nabla_w f(u,w)), \quad x = (u,w), \quad Q = Q_u \otimes Q_w,$$

где функция $f(u,w)$ выпуклая по u и вогнутая по w, то из неравенства (5.4) следует, что

$$0 \leqslant \max_{w\in Q_w} f(\overline{u}^N, w) - \min_{u\in Q_u} f(u, \overline{w}^N) \leqslant \varepsilon.$$

Отметим при этом, что для седловой точки (u_*, w_*) выполнено равенство

$$\max_{w\in Q_w} f(u_*, w) = \min_{u\in Q_u} f(u, w_*). \ \blacklozenge$$

Заметим, что для *монотонных вариационных неравенств*[4]

$$\langle g(y) - g(x), y - x\rangle \geqslant 0, \quad x, y \in Q,$$

[3] Если это условие не выполняется, то решение монотонного ВН можно свести к задаче негладкой выпуклой оптимизации, которую можно решать методами типа центров тяжести [52, 74], см. указание к упражнению 1.4. Отметим также, что метод эллипсоидов можно применять для решения ВН и непосредственно [465].

[4] В случае $g(x) = \nabla f(x)$ это условие соответствует выпуклости функции $f(x)$.

выполнено условие

$$\langle g(x), y^k - x\rangle = \langle g(y^k), y^k - x\rangle + \underbrace{\langle g(x) - g(y^k), y^k - x\rangle}_{\leqslant 0} \leqslant \langle g(y^k), y^k - x\rangle.$$

В этой связи формулу (5.4) можно переписать как[5]

$$\max_{x\in Q}\langle g(x), \bar{y}^N - x\rangle \leqslant \varepsilon, \tag{5.6}$$

где

$$\bar{y}^N = \frac{1}{\sum_{k=1}^N \frac{1}{L^k}} \sum_{k=1}^N \frac{y^k}{L^k}.$$

В литературе обычно используют именно этот критерий качества решения монотонных ВН, см., например, [75, гл. 3], [466].

Подобно слабой квазивыпуклости из замечания 2.1, можно ослабить условия монотонности ВН, во многом сохранив результаты работы [223]. ■

Замечание 5.2 (негладкие задачи и рандомизированные методы). Как уже отмечалось, основным достоинством универсального подхода является автоматическая и адаптивная настройка на гладкость задачи. И даже если задача заведомо негладкая, универсальный подход может давать существенные преимущества по сравнению с оптимальными методами, настроенными на негладкие задачи, см., например, [7]. Однако у универсального подхода есть несколько минусов. Во-первых, это не адаптивный подход в том смысле, что в метод явным образом зашита желаемая точность ε (к чему это приводит, см., например, в формуле (2.34))[6]. Отказавшись от универсальности, можно избавиться и от этого ограничения, используя *метод двой-*

[5] Собственно под слабым решением вариационного неравенства обычно понимают следующую задачу [149]: найти такую точку $x_* \in Q$, что для всех $x \in Q$ выполнено неравенство $\langle g(x), x_* - x\rangle \leqslant 0$, т. е. $\max_{x\in Q}\langle g(x), x_* - x\rangle \leqslant 0$. Отсюда становится ясным смысл условия (5.6). Подробнее о ВН см., например, [6, 35, 62, 389, 545], [52, часть 3], [75, гл. 3]. В частности, полезно заметить, что ВН можно решать и в сильном смысле [149]: найти такую точку $x_* \in Q$, что для всех $x \in Q$ выполнено неравенство $\langle g(x_*), x_* - x\rangle \leqslant 0$, т. е. $\max_{x\in Q}\langle g(x_*), x_* - x\rangle \leqslant 0$. Если векторное поле $g(x)$ непрерывное и монотонное (ВН монотонное), то слабые и сильные решения ВН совпадают [52, часть 3, лемма 2.2.1] и можно говорить просто о решении ВН. В общем же случае сильное решение всегда будет слабым, но не наоборот.

[6] Впрочем, адаптивность по желаемой точности ε можно получить, рестартуя универсальный метод с новым значением точности $\varepsilon := \varepsilon/2$ в момент, когда достигнута точность ε, и т. д.

ственных усреднений Ю. Е. Нестерова[7] [481]. Во-вторых, обоснование метода требует возможности проведения выкладок хотя бы в общности § 2, однако в действительности для негладких задач достаточно общности (1.32), что заметно упрощает и вывод основных оценок, и обоснование возможности последующего обобщения на стохастические постановки [96]. Отметим тем не менее, что здесь речь идёт только о простоте выводов, но не о потенциальных возможностях вывода. В-третьих, привязка к соотношению (1.32) позволяет переносить результаты, полученные непосредственно для негладких выпуклых задач, т. е. без универсализации, на онлайн-постановки, в том числе стохастические и сильно выпуклые[8], см., например, [334, 500]. В-четвёртых, для негладких задач концепция неточного оракула (см., например, (2.3), (3.1)) может быть заменена на более простую и менее ограничительную концепцию δ-субградиента (см., например, [86, п. 5, § 1 и п. 3, § 3, гл. 5]), в которой отсутствуют правые неравенства в формулах (2.3), (3.1)[9]. В-пятых, при перенесении универсальных методов на стохастические постановки задач [31], в которых случайность искусственно ввели мы сами (это, как правило, называется *рандомизацией* метода) при вычислении градиента или проектировании, чтобы сократить вычислительную сложность этих операций взамен на увеличение их числа, из-за погони за универсальностью могут теряться некоторые свойства дешевизны этих операций [32]. Связано это прежде всего с тем, что при универсальном подходе необходимо рассчитывать значение функции, что может быть намного дороже расчёта значения её стохастического градиента. Вот простой пример [366]:

$$f(x) = \frac{1}{2}\langle x, Ax\rangle,$$

[7] Этот метод близок другому популярному методу решения негладких задач оптимизации — *методу зеркального спуска* А. С. Немировского, см., например, [127], [164, гл. 5], [365] и упражнение 2.6.

[8] Отметим, что конструкция рестартов (упражнение 2.3) в онлайн-постановках уже не работает. Более того, для сильно выпуклых задач онлайн-оптимизации оценка (2.37) уже не достижима. Необходима её корректировка в части

$$\frac{L_0^2}{\mu N} \to \frac{L_0^2 \ln N}{\mu N}.$$

Такая нижняя оценка уже будет достижима [334, 335].

[9] Работая с δ-субградиентами где $\delta = O(\varepsilon)$ [301], вместо настоящих субградиентов можно получать, например, оценки из указания к упражнению 1.4, изначально установленные для работы с настоящими субградиентами [74, 135], см. также упражнение 5.5.

где $A \succ 0$ — плотно заполненная неотрицательно определённая матрица размера $n \times n$, $x \in S_n(1)$. Несмещённый стохастический градиент этой функции имеет вид

$$\nabla_x f(x, j) = A^j, \quad \text{где } P(j = i) = x_i,\ i = 1, \ldots, n.$$

Ясно, что

$$E_j[\nabla_x f(x, j)] = Ax = \nabla f(x).$$

Также ясно, что на подсчёт $f(x)$ уходит время $O(n^2)$, в то время как на подсчёт $\nabla_x f(x, j)$ — время $O(n)$. Впрочем, если для решения задачи используется *минибатчинг*, что более характерно для решения задач стохастической оптимизации, чем при рандомизации детерминированных процедур (см. приложение), то такой проблемы не возникает. ■

♦ Вернёмся к задаче минимизации квадратичной формы на симплексе из замечания 5.2 и упражнения 1.6, считая, что матрица A, задающая квадратичную форму, плотно заполненная и все элементы этой матрицы ограничены по модулю числом M: $|A_{ij}| \leqslant M$. Если для решения этой задачи использовать быстрый градиентный метод, то оценка общего времени работы метода, необходимого для достижения точности по функции ε, будет следующей:

$$\underbrace{O(n^2)}_{\substack{\text{сложность}\\ \text{итерации}}} \cdot \underbrace{O\left(\sqrt{\frac{L_1 R^2}{\varepsilon}}\right)}_{\substack{\text{число}\\ \text{итераций}}} = O\left(n^2 \sqrt{\frac{M \ln n}{\varepsilon}}\right).$$

В этом примере используется 1-норма и энтропия в качестве прокс-функции, см. конец § 2 и упражнение 3.7. Если ту же самую задачу решать с той же точностью ε (только в среднем) рандомизированным методом с рандомизацией, описанной в замечании 5.2, то оценка общего времени работы будет составлять

$$\underbrace{O(n)}_{\substack{\text{сложность}\\ \text{итерации}}} \cdot \underbrace{O\left(\frac{L_0^2 R^2}{\varepsilon^2}\right)}_{\substack{\text{число}\\ \text{итераций}}} = O\left(n \cdot \frac{M^2 \ln n}{\varepsilon^2}\right).$$

Для задач очень больших размеров при невысоких требованиях к точности решения второй (рандомизированный) способ может оказаться предпочтительнее. ♦

Вернёмся к оценке (5.1). Попробуем с помощью этой оценки и *техники рестартов* (упражнение 2.3) распространить описанный выше

универсальный градиентный спуск на сильно выпуклые задачи. Ввиду теоремы 1.1 отметим, что существуют и другие способы, позволяющие это сделать. Однако выбранный здесь способ представляется наиболее удачным в методическом плане своей общепримени мостью.

Итак, подобно выводу неравенства (4.5) из (4.1), с помощью неравенства (5.1) можно получить оценку

$$f(\bar{x}^N) - f(x_*) \leqslant \frac{\bar{L}_N V(x_*, x^0)}{2N} + \frac{\varepsilon}{2}, \tag{5.7}$$

где

$$\bar{L}_N = \frac{N}{\sum_{k=1}^N \frac{1}{L^k}}, \quad \bar{x}^N = \frac{\bar{L}_N}{N} \sum_{k=1}^N \frac{x^k}{L^k}.$$

Пусть $f(x)$ — μ-сильно выпуклая функция в норме $\|\ \|$, согласованной с дивергенцией Брэгмана $V(y, x)$ (см. § 2, в частности формулу (2.32)). Пусть

$$d(x - x^0) \leqslant C_n \|x - x^0\|^2$$

(можно считать, что $C_n = \mathrm{O}(\ln n)$, см. п. 2 упражнения 2.3). Тогда из неравенств (1.14) и (5.7) следует, что

$$\frac{\mu}{2}\|\bar{x}^{N_1} - x_*\|^2 \leqslant f(\bar{x}^{N_1}) - f(x_*) \leqslant \frac{\bar{L}_{N_1} V(x_*, x^0)}{2N_1} + \frac{\varepsilon}{2} \leqslant \frac{C_n \bar{L}_{N_1} \|x^0 - x_*\|^2}{2N_1} + \frac{\varepsilon}{2}.$$

Далее используется схема рассуждений, аналогичная той, что была изложена в указании к упражнению 2.3. А именно, из соотношения

$$\frac{\mu}{2}\|\bar{x}^{N_1} - x_*\|^2 \leqslant \frac{C_n \bar{L}_{N_1} \|x^0 - x_*\|^2}{2N_1} + \frac{\varepsilon}{2} \tag{5.8}$$

выбираем наименьшее такое N_1, при котором

$$\|\bar{x}^{N_1} - x_*\|^2 \leqslant \frac{1}{2}\|x^0 - x_*\|^2. \tag{5.9}$$

Для этого не надо знать x_*, можно просто воспользоваться соотношением (5.8). Однако при этом необходимо знать μ. В итоге можно показать, что для такого метода[10] из оценки (5.2) получится оценка

[10] На $(k+1)$-м рестарте следует выбирать точку старта как $x^0 = \bar{x}^{N_k}$ — среднее арифметическое точек, полученных по ходу работы метода на k-м рестарте, а $d(x) := d(x - x^0) = d(x - \bar{x}^{N_k})$ [31, 366, 369]. Натуральный логарифм в формуле (5.10) выбран потому, что именно такой выбор наиболее часто встречается в данном контексте в литературе. Отметим также, что подписи сомножителей в формуле (5.10) (и далее в аналогичных формулах) приводятся с точностью до числовых множителей.

вида [19, 27, 269]

$$N = \mathrm{O}\Bigg(\underbrace{C_n \inf_{\nu\in[0,1]} \Big(\frac{L_\nu^2}{\mu^{1+\nu}\varepsilon^{1-\nu}}\Big)^{1/(1+\nu)}}_{\substack{\text{число итераций}\\ \text{на одном рестарте}}} \underbrace{\Big\lceil \ln\Big(\frac{\mu\|x_* - x^0\|^2}{\varepsilon}\Big)\Big\rceil}_{\text{число рестартов}} \Bigg); \tag{5.10}$$

здесь и далее $\lceil a \rceil = \max\{1, a\}$. Заметим, что оценки (1.24), (2.37) соответствуют (5.10) с точностью до логарифмического множителя при $\nu = 1$, $\nu = 0$ соответственно.

Формулу (5.10) можно проверить с помощью регуляризации (см. замечание 4.1). А именно, исходную выпуклую задачу всегда можно сделать сильно выпуклой с константой сильной выпуклости $\mu \simeq \varepsilon/R^2$. Подставляя $\mu \simeq \varepsilon/R^2$ в формулу (5.10) с точностью до C_n, получим оценку (5.2).

♦ В работе [369] (см. также [31, 75]) отмечается, что число обусловленности (отношение константы Липшица градиента L^1 к константе сильной выпуклости μ^1), например, для квадратичных функций

$$f(x) = \frac{1}{2}\langle x, Ax\rangle - \langle b, x\rangle, \quad x \in \mathbb{R}^n,$$

посчитанное в 1-норме, не может быть меньше n, в то время как число обусловленности, посчитанное в евклидовой норме (2-норме), может при этом равняться 1. Действительно, пусть $\xi = (\xi_1, \ldots, \xi_n)$, где ξ_k — независимые одинаково распределённые случайные величины,

$$P\Big(\xi_k = \frac{1}{n}\Big) = P\Big(\xi_k = -\frac{1}{n}\Big) = \frac{1}{2}.$$

Тогда, учитывая, что $\|\xi\|_1 \equiv 1$, получим

$$\mu^1(f) \leqslant E_\xi[\xi^T A \xi] = \frac{1}{n^2}\mathrm{tr}(A) \leqslant \frac{1}{n}\max_{i,j=1,\ldots,n}|A_{ij}| = \frac{1}{n}L^1(f).$$

Отсюда напрашивается вывод, что при решении сильно выпуклых задач естественно выбирать евклидову норму. Как правило, это действительно так. Однако, во-первых, отмеченное выше наблюдение совсем не означает, что обусловленность задачи в 1-норме всегда хуже, чем в 2-норме. Например, числа обусловленности энтропии $f(x) = \sum_{i=1}^n x_i \ln x_i$ на множестве

$$\Big\{x \in \mathbb{R}^n \colon \sum_{i=1}^n x_i = 1;\ x_i \geqslant \delta,\ i = 1, \ldots, n\Big\}$$

в 1-норме и 2-норме совпадают и равны $1/\delta \geqslant n$. Во-вторых, для задач композитной оптимизации [471] (см. также пример 3.1) L^1 можно брать только от гладкого слагаемого, а μ^1 от всей функции. Таким образом можно решать, например, задачу [27]

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2 + \mu \sum_{i=1}^{n} x_i \ln x_i \to \min_{x \in S_n(1)}. \quad \blacklozenge$$

Основная проблема с реализацией описанного выше подхода — это явное использование в нём, как правило, неизвестного параметра μ. Довольно естественный способ борьбы с неизвестностью параметра μ состоит в рестартах по невязке $f(x^N) - f(x_*)$. Если $f(x_*)$ известно, то рестарты можно делать, контролируя эту невязку по функции [269, 292, 532]. Отметим, что если известно $f(x_*)$, то можно также предложить адаптивный по μ (ускоренный) градиентный метод, не требующий рестартов [153].

Замечание 5.3 (контроль нормы градиента). Для неускоренного (универсального) градиентного спуска также можно использовать рестарты по норме градиента (градиентного отображения; см. упражнение 3.4)[11]. Далее мы для наглядности ограничимся задачей безусловной выпуклой оптимизации (1.1) в условиях (1.4). Тогда для метода (1.22)

$$x^{k+1} = x^k - \frac{1}{L}\nabla f(x^k)$$

имеет место следующая не улучшаемая для данного метода оценка [582] (следует сравнить с оценкой (1.23), оптимальной для класса гладких невыпуклых задач):

$$\|\nabla f(x^N)\|_2 \leqslant \frac{LR}{N+1},$$

где $R = \|x^0 - x_*\|_2$. Несложно показать, что рестарты (на базе формул (1.14), (1.15)) с данной оценкой приводят к правильным порядкам скорости сходимости (в том числе по функции и по аргументу) для неускоренных методов в сильно выпуклом случае (1.24) [582]. Ситуация меняется для ускоренных методов, см. указание к упражнению 1.3. Для ускоренных (быстрых, моментных) градиентных спусков на данный момент удаётся получить (см. неравенство (1.8)) лишь оценки

[11] Следует сопоставить это с тем, что ранее отмечалось в комментариях к упражнению 2.3 относительно невозможности использовать такой критерий для рестартов в общем случае (например, для ускоренных методов).

вида[12] [386, 472, 582]

$$\|\nabla f(x^N)\|_2 \leqslant \sqrt{2L\cdot(f(x^N)-f(x_*))} \leqslant \frac{2LR}{N+1}, \quad \min_{k=0,\dots,N}\|\nabla f(x^k)\|_2 \leqslant \frac{8LR}{N^{3/2}},$$

которые не улучшаемы больше чем на числовой множитель для рассмотренных на данный момент классов (ускоренных) методов первого порядка [582]. До недавнего времени открытым оставался вопрос, можно ли добиться для какого-нибудь из методов первого порядка сходимости вида $\|\nabla f(x^N)\|_2 \sim LR/N^2$ [472, 582]. Для такого метода можно было бы делать рестарты по норме градиента без риска потерять оптимальность. Тем не менее если регуляризовать исходную задачу (см. замечание 4.1) и решать регуляризованную задачу любым вариантом быстрого градиентного метода, уже настроенного на сильно выпуклую постановку [31, 75, 76, 186, 236], то с точностью до $\ln N$ получается желаемая оценка[13]. К сожалению, для всех известных сейчас вариантов таких методов в размер шага явно входит константа сильной выпуклости, неизвестная в данном контексте, см. указание к упражнению 1.3. К тому же не совсем понятно, какой смысл бороться здесь за оптимальную оценку, чтобы предложить на её основе с помощью рестартов оптимальный метод для гладких сильно выпуклых задач, если получается, что вся эта конструкция, в свою очередь, сама базируется на таком методе.

В 2018 г. в работе [388] был предложен такой вариант быстрого градиентного метода, для которого

$$\|\nabla f(x^N)\|_2^2 = \mathrm{O}\Big(\frac{L\cdot(f(x^0)-f(x_*))}{N^2}\Big).$$

Излагаемая далее конструкция (два абзаца текста) была нам сообщена Ю. Е. Нестеровым. Если сначала запустить обычный быстрый градиентный метод на N итераций, то получим

$$f(x^N)-f(x_*) = \mathrm{O}\Big(\frac{LR^2}{N^2}\Big).$$

[12] Последняя оценка получается для сочетания сначала $N/2$ шагов быстрого градиентного метода, затем $N/2$ шагов обычного градиентного метода [472].

[13] Упражнение 4.2 и данное предложение проясняют необходимость использования регуляризации при решении двойственной задачи ускоренными методами (см. упражнение 4.4) в качестве альтернативы к использованию прямодвойственных методов. Двойственная задача регуляризуется, чтобы оптимально (с точностью до логарифмических множителей) восстанавливать по найденному приближённому решению двойственной задачи решения прямой задачи [24, 235, 266].

Запустив затем метод из работы [388], получим

$$\|\nabla f(x^{2N})\|_2^2 = \mathrm{O}\Big(\frac{L\cdot(f(x^N)-f(x_*))}{N^2}\Big) = \mathrm{O}\Big(\frac{L^2R^2}{N^4}\Big).$$

Приведённые оценки хорошо согласуются с результатами работы [463] (см. также замечание 1.6) и в общем случае неулучшаемы.

Отметим, однако, одну сложность, не позволяющую для метода из работы [388] использовать норму градиента в качестве критерия останова. Методу из работы [388] требуется давать на вход в качестве одного из параметров число итераций N, которые метод должен сделать. Таким образом, приведённая выше оценка скорости сходимости с параметром N не предполагает, что метод не зависит от этого параметра. Для каждого значения N получается свой метод. Тем не менее такой метод можно использовать с конструкцией рестартов при известном значении L, но неизвестном значении μ более эффективно, чем обычный быстрый градиентный метод. Для этого нужно брать в качестве критерия окончания рестарта условие уполовинивания нормы градиента на этом рестарте, а в качестве числа итераций на k-м рестарте — $N^k = \sqrt{L/\mu^k}$, где $\mu^k = 2\mu^{k-1}$, и далее подбирать параметр μ^k с возможным понижением: $\mu^k := \mu^{k-1}/2$, до тех пор пока не выполнится критерий окончания рестарта [510].

Отметим, что с помощью техники, аналогичной технике, использованной в работе [388] (см. также замечание 1.5 и последующий поясняющий текст), для выпукло-вогнутых седловых задач (см. замечание 5.2) с L-липшицевым градиентом

$$\min_u \max_w f(u,w)$$

пока можно получить лишь оценку на число вычислений $\nabla_u f(u,w)$ и $\nabla_w f(u,w)$ [384]:

$$\|\nabla f(u^N,w^N)\|_2^2 = \widetilde{\mathrm{O}}\Big(\frac{L^2R^2}{N^2}\Big).$$

Можно ли здесь убрать логарифмический множитель (заменить $\widetilde{\mathrm{O}}(\,)$ на $\mathrm{O}(\,)$)? Ответ на этот вопрос, насколько мы знаем, по-прежнему является открытым. Отметим также, что конструкция из упомянутой работы [384], по-видимому, позволяет в некотором смысле перенести идею каталиста на седловые задачи. ■

Для задач безусловной выпуклой оптимизации замечание 5.3 также даёт возможность контролировать только норму градиента в качестве критерия останова неускоренного градиентного спуска и его универсального варианта.

Отметим также, что в работе [280] недавно была обнаружена возможность (не связанная с использованием прямодвойственности) избавления от знания значения параметра μ и в рестартах для ускоренных вариантов градиентных методов.

В заключение обратим внимание на то, что для задач оптимизации на неограниченных множествах не могут одновременно выполняться оба используемых выше неравенства: $\mu > 0$ и $L_\nu < \infty$, где $\nu \in [0, 1)$. Однако ввиду неравенств типа (3.16) можно быть уверенным в том, что последовательность точек, сгенерированных методом универсального градиентного спуска с рестартами, будет лежать в компактном множестве, задаваемом формулой (3.16), на котором и стоит определять константы $\mu > 0$ и $L_\nu < \infty$, входящие в итоговую оценку числа итераций (5.10).

Упражнение 5.1. Предложите универсальный градиентный спуск в общности § 3, т. е. используя общую концепцию модели функции в точке. Покажите, что оценки (5.2), (5.5) сохранят свой вид, если допускать неточности $\delta = O(\varepsilon)$ и $\tilde{\delta} = O(\varepsilon)$ (см. обозначения из § 3). Поясните, как следует понимать эти неточности для подхода из замечания 5.1, приводящего к оценке (5.5).

Упражнение 5.2. Попробуйте сформулировать и доказать утверждение, аналогичное утверждению из упражнения 2.2, для универсального градиентного спуска.

Упражнение 5.3. Попробуйте распространить упражнение 5.1 на сильно выпуклые постановки задач.

Указание. См. [568,569]. Оптимизируемая (целевая) функция $f(x)$, как и раньше, предполагается выпуклой на выпуклом множестве Q. Условие гладкости и сильной выпуклости функции $f(x)$ при наличии шума следует понимать следующим образом (3.1) [236, 271]:

$$\frac{\mu}{2}\|y-x\|^2 \leqslant f(y) - \left(f_\delta(x) + \psi_\delta(y,x)\right) \leqslant \frac{L}{2}\|y-x\|^2 + \delta.$$

Впрочем, используя технику рестартов, можно исходить и из обычной концепции модели функции (3.1), предполагая сильную выпуклость у модели функции $\psi_\delta(y,x)$ (см. формулу (3.1)) как функции переменной y или (ещё более общий случай) предполагая сильную выпуклость только у $f(x)$ [31, 269].

Упражнение 5.4. С помощью техники рестартов (см. упражнение 2.3) и формулы (5.3) (или иным способом [91]) попробуйте перенести результаты замечания 5.1 на *сильно монотонные вариационные*

неравенства и сильно выпукло-вогнутые седловые задачи [75, гл. 3]. Отметим одно затрудняющее такой перенос обстоятельство: в формуле (5.5) используется $R^2 = \max_{x\in Q} V(x, x^0)$, что не даёт возможности формально применять технику рестартов. Можно ли получить для универсального проксимального зеркального метода, использующегося для решения ВН и седловых задач, оценки, подобные (2.19)?

Указание. Для наглядности будем считать, что $V(x, y) = \|x-y\|_2^2/2$. Обозначим через x_* решение ВН. Заметим, что для всех $y^k \in Q$ выполнено неравенство $\langle g(x_*), y^k - x_*\rangle \geqslant 0$. Из формулы (5.3) следует, что

$$\frac{1}{N}\sum_{k=1}^{N}\langle g(y^k), y^k - x_*\rangle \leqslant \frac{L\|x_* - x^0\|_2^2}{2N}.$$

Объединяя эти два неравенства, получим

$$\frac{1}{N}\sum_{k=1}^{N}\langle g(y^k) - g(x_*), y^k - x_*\rangle \leqslant \frac{L\|x_* - x^0\|_2^2}{2N}.$$

По предположению ВН сильно монотонное, т. е. существует такое $\mu > 0$, что

$$\langle g(y) - g(x), y - x\rangle \geqslant \mu\|y - x\|_2^2, \quad x, y \in Q.$$

Учитывая это неравенство и выпуклость функции $\|x\|_2^2$, получим

$$\mu\|\bar{y}^N - x_*\|_2^2 \leqslant \mu\sum_{k=1}^{N}\frac{1}{N}\|y^k - x_*\|_2^2 \leqslant$$
$$\leqslant \frac{1}{N}\sum_{k=1}^{N}\langle g(y^k) - g(x_*), y^k - x_*\rangle \leqslant \frac{L\|x_* - x^0\|_2^2}{2N},$$

где

$$\bar{y}^N = \frac{1}{N}\sum_{k=1}^{N} y^k.$$

На основе последнего неравенства уже можно организовать процедуру рестартов, получив следующую оценку на общее (суммарное) число итераций:

$$O\Big(\frac{L}{\mu}\ln\Big(\frac{\mu R^2}{\varepsilon}\Big)\Big).$$

Данная оценка в общем случае уже не может быть улучшена для рассматриваемого класса ВН никакими другими методами [502, 616]. Заметим также, что приведённую оценку можно получить и в модельной общности [568].

Упражнение 5.5. Используя принцип множителей Лагранжа [182, гл. 5], распространите проксимальный зеркальный метод из замечания 5.1 на общие задачи выпуклого программирования [66, п. 2], предварительно компактифицировав двойственные переменные (см., например, упражнение 4.1 и замечание 4.2). Рассмотрите альтернативный подход к решению возникшей седловой задачи в случае, когда общее число аффинных ограничений вида равенств и выпуклых ограничений вида неравенств мало. В качестве альтернативного подхода предлагается решать двойственную задачу прямодвойственным методом эллипсоидов [219, 465]. Для этого потребуется число итераций $N_{\text{ellips}}(\varepsilon) \sim \ln \varepsilon^{-1}$ (см. указание к упражнению 1.4). При этом на каждой итерации вместо настоящего субградиента можно использовать δ-субградиент (см. замечание 5.2), где $\delta \sim \varepsilon$, который вычисляется по формуле Демьянова — Данскина из решения с относительной точностью (по функции) δ вспомогательной задачи выпуклой оптимизации, получающейся из рассматриваемой седловой задачи при фиксации двойственных переменных [86, п. 5, § 1, гл. 5]. Проведите описанные выше рассуждения более аккуратно, прорабатывая детали.

Указание. Следует сопоставить данное упражнение с упражнениями 4.3, 4.6, 5.6 и примером 3.2.

Упражнение 5.6. Предложите способ решения задачи минимизации достаточно гладкой выпуклой функции при наличии, вообще говоря, негладкого скалярного сильно выпуклого ограничения вида неравенства. При этом решение задачи должно обращать это неравенство в равенство.

Указание. Следует воспользоваться упражнением 5.5. При этом на каждой итерации метода в двойственном пространстве на вспомогательную задачу оптимизации следует смотреть как на задачу композитной сильно выпуклой оптимизации (см. пример 3.1), чтобы негладкость ограничения не учитывалась, а его сильная выпуклость, напротив, позволяла решать вспомогательную задачу за линейное время, используя, например, технику рестартов, см. конец § 5 и работу [27].

Отметим, что если в условии задачи имеется несколько сильно выпуклых ограничений вида неравенств $h_1(x) \leqslant 0, \ldots, h_m(x) \leqslant 0$, то их можно заменить скалярным негладким сильно выпуклым ограничением:

$$h(x) = \max\{h_1(x), \ldots, h_m(x)\} \leqslant 0$$

— см. [3], [86, п. 3, § 3, гл. 10], [484, 491] и замечание 3.1.

Заметим, что если функция, задающая сильно выпуклое ограничение, гладкая (имеет липшицев градиент), то нет необходимости её считать композитно-дружественной. При этом на вспомогательную задачу можно смотреть как на задачу обычной (не композитной) оптимизации.

Упражнение 5.7. Предложите способ распространения проксимального зеркального метода из замечания 5.1 на бесконечномерные задачи, например дифференциальные игры [275].

Упражнение 5.8. Определите, какие из результатов, описанных выше (во всём пособии), могут быть перенесены с обычного (неускоренного) градиентного метода на ускоренные (быстрые, моментные) градиентные методы.

Указание. На метод подобных треугольников из упражнения 3.7 переносятся все результаты, кроме результатов, собранных в замечаниях 1.1, 1.3, 5.1, 5.3, и результатов, для которых лучше подходит ускоренный проксимальный метод Монтейро — Свайтера или его аналоги, см. замечания 1.6, 3.3. В случае замечания 5.1 частичное ускорение при некоторых дополнительных предположениях оказывается возможным [2, 26, 207, 208, 255, 435, 596].

Отметим, что результаты из § 3, связанные с относительной гладкостью, переносятся на ускоренные методы лишь при дополнительных обременительных предположениях на дивергенцию Брэгмана [246, 328, 330, 343, 439]. По-видимому, аналогичные сложности не позволяют предложить полноценный проксимальный ускоренный градиентный метод с неевклидовым проксом, см. замечание 3.3 и текст сразу после него.

Результаты, касающиеся α-слабой квазивыпуклости, переносятся на ускоренные методы [324, 345, 470].

Упражнение 5.9 (Ю. Е. Нестеров, 2014). Задача поиска такого x_*, что $Ax_* = b$, сводится к задаче выпуклой гладкой оптимизации:

$$f(x) = \|Ax - b\|_2^2 \to \min_{x \in \mathbb{R}^n}.$$

Нижняя оценка при $N \leqslant n$ на скорость решения такой задачи (см. также упражнение 1.3 и замечание 1.6) имеет вид

$$\|Ax^N - b\|_2^2 \geqslant \frac{L_x R_x^2}{2(2N+1)^2},$$

где $L_x = \sigma_{\max}(A)$, $R_x = \|x_*\|_2$. Если решение x_* не единственное, то в определении R_x можно считать, что используется решение с наи-

меньшей 2-нормой. При этом на каждой итерации разрешено не более двух раз умножать матрицу A на вектор (справа и слева). С другой стороны, рассмотрим задачу

$$\frac{1}{2}\|x\|_2^2 \to \min_{Ax=b}.$$

Построим к ней двойственную задачу [182, гл. 5]:

$$\min_{Ax=b}\frac{1}{2}\|x\|_2^2 = \min_x \max_\lambda \left\{\frac{1}{2}\|x\|_2^2 + \langle b - Ax, \lambda\rangle\right\} =$$
$$= \max_\lambda \min_x \left\{\frac{1}{2}\|x\|_2^2 + \langle b - Ax, \lambda\rangle\right\} = \max_\lambda \left\{\langle b, \lambda\rangle - \frac{1}{2}\|A^T\lambda\|_2^2\right\}.$$

С помощью теоремы 4.1 и упражнений 1.3, 5.8 (альтернативный способ базируется на замечании 5.3) покажите, что после N итераций ускоренного градиентного метода, применённого к двойственной задаче, можно восстановить решение исходной задачи $\check{x}^N$ со следующей точностью:

$$\|A\check{x}^N - b\|_2 \leqslant \frac{8L_\lambda R_\lambda}{N^2},$$

где $L_\lambda = \sigma_{\max}(A^T) = \sigma_{\max}(A)$, $R_\lambda = \|\lambda_*\|_2$. Если решение λ_* не единственное, то в определении R_λ можно считать, что используется решение с наименьшей 2-нормой. При этом общее число умножений матрицы A на вектор не будет превышать $2N$. Поясните, почему последняя оценка не противоречит выписанной ранее нижней оценке.

Указание. См. [4, 20, 143, 292, 524] и замечание 1.6. При этом важно заметить, что связь оптимальных значений в прямой и двойственной задаче $x_* = A^T\lambda_*$, которая следует из связи прямых и двойственных переменных $x(\lambda) = A^T\lambda$, и условие $R_\lambda = \|\lambda_*\|_2$ можно понимать в совокупности как *условие истокопредставимости*, см. также упражнение 4.9.

Заметим также, что, поскольку система $Ax = b$ совместна, по *теореме Фредгольма* [52, п. 2.6. Ч. 1] не существует такого λ, что $A^T\lambda = 0$ и $\langle b, \lambda\rangle > 0$, следовательно, двойственная задача имеет конечное решение, т. е. существует ограниченное решение двойственной задачи λ_*. Действительно, по предположению существует такой вектор x, что $Ax = b$, поэтому для всех λ имеет место равенство $\langle Ax, \lambda\rangle = \langle b, \lambda\rangle$. Следовательно, $\langle x, A^T\lambda\rangle = \langle b, \lambda\rangle$. Предположив, что существует такой вектор λ, что $A^T\lambda = 0$ и $\langle b, \lambda\rangle > 0$, придём к противоречию: $0 = \langle x, A^T\lambda\rangle = \langle b, \lambda\rangle > 0$.

♦ В замечании 1.5 отмечалось, что решение системы линейных уравнений $Ax = b$ является краеугольным камнем не только (вычислительной) линейной алгебры [304, 499, 536, 585] и вычислительной матема-

тики [95], но и численных методов оптимизации [183]. Напомним также, что класс сложности задач выпуклой безусловной оптимизации в категориях O( ) характеризуется классом сложности задач квадратичной оптимизации (см. упражнение 1.3 и замечание 1.6) и что необходимость в решении системы линейных уравнений (обращении матрицы) возникает на каждом шаге метода Ньютона (см. приложение).

В свою очередь, задачи квадратичной оптимизации обычно получаются из задачи $Ax = b$ либо как указано в упражнении 5.9, либо (в случае симметричности матрицы A) согласно соотношению (1.30), см. также замечание 1.6. Упражнения 1.6, 5.9 (см. также [32, гл. 4]) показывают, что в случае дополнительных предположений можно пытаться решать (разреженные) линейные системы быстрее, чем предписывают нижние оценки, полученные из нижних оценок для задач квадратичной оптимизации. Предполагая выполненными некоторые свойства матрицы[14] A, также можно решать (используя рандомизированные методы) системы линейных уравнений за время, пропорциональное $m = nnz(A)$ (с точностью до больших степеней логарифмических множителей, зависящих от желаемой точности) — числу ненулевых элементов в матрице A [32, гл. 4], [213, 380, 525, 562]. Отметим, что в основе части отмеченных работ лежит полезное наблюдение [562, 593]: для матрицы Лапласа можно построить за время $\widetilde{O}(m)$ другую матрицу Лапласа (на основе остовного дерева неориентированного графа исходной матрицы с добавлением ещё нескольких рёбер — полученный граф называют *ультра-спарсификатором*), которая:

1) хорошо предобуславливает (см., например, [499]) исходную матрицу A; обобщённое число обусловленности[15] [166] становится равным $\widetilde{O}(1)$;

[14] Например, симметричность и слабое диагональное доминирование. Это свойство выполняется, в частности, для матрицы Лапласа неориентированного графа, см. пример 4.1.

[15] Чтобы приблизительно понять, чем обобщённое число обусловленности отличается от обычного числа обусловленности, вернёмся к комментарию к упражнению 4.7, в котором рассматривалась матрица Лапласа графа со звёздной топологией. Было показано, что если эту матрицу умножить на специальную диагональную матрицу, то полученная в результате (уже не симметричная) матрица будет иметь неотрицательный действительный спектр с отношением максимального собственного значения к минимальному ненулевому, равным приблизительно 2, в то время как отношение соответствующих (максимального и минимального ненулевого) сингулярных чисел этой матрицы будет иметь порядок n. Обобщённое число обусловленности здесь равно 2, а не n.

2) содержит $\widetilde{\mathrm{O}}(m)$ ненулевых элементов;
3) эффективно обратима.

Далее, итеративно сочетая отмеченное предобуславливание задачи с итерациями метода сопряжённых градиентов (см. замечание 1.6), можно получить отмеченный выше результат. Заметим, что метод сопряжённых градиентов, гарантированно сходящийся к точному решению в общем случае (без предобуславливания) за n итераций, на каждой своей итерации требует умножения матрицы A на вектор. Таким образом, сложность только одной такой итерации пропорциональна числу ненулевых элементов матрицы A. Интересно сравнить отмеченную (как правило, завышенную) оценку сложности метода сопряжённых градиентов $\mathrm{O}(mn)$ с оценкой, основанной на быстром матричном умножении $\widetilde{\mathrm{O}}(n^{2{,}37})$ (см. также конец приложения), и оценкой $\widetilde{\mathrm{O}}(m)$ [562]. Тем не менее в смысле универсальности, лёгкой имплементируемости и реальной скорости работы на практике обычный метод сопряжённых градиентов является, как правило, более предпочтительным. На данный момент отмеченная технология [562] представляет интерес в большей степени с теоретической точки зрения ввиду возникновения больших степеней у возникающих логарифмических множителей в оценках.

В связи с изучением задач децентрализованной распределённой оптимизации (см. пример 4.1) описанная выше матрица Лапласа ультра-спарсификатора исходного графа (сети) позволяет разрабатывать эффективные конструкции разреженных и хорошо обусловленных коммуникационных сетей. ♦

Приложение. Обзор современного состояния развития численных методов выпуклой оптимизации

Описанные в § 2–5 конструкции переносятся на ускоренные (быстрые, моментные) градиентные методы [97], например на *метод подобных треугольников*[1] [127], см. также упражнения 3.7, 5.8. При этом дальнейшее ускорение в общем случае уже невозможно (см. упражнение 1.3). Для ускоренного метода оценки (5.2), (5.10) и условия на допустимый уровень шума δ преобразуются следующим образом [19, 27, 71, 97, 269, 492]:

$$N(\varepsilon) = \mathrm{O}\left(\inf_{\nu\in[0,1]}\left(\frac{L_\nu R^{1+\nu}}{\varepsilon}\right)^{2/(1+\nu)}\right) \to \mathrm{O}\left(\inf_{\nu\in[0,1]}\left(\frac{L_\nu R^{1+\nu}}{\varepsilon}\right)^{2/(1+3\nu)}\right),$$

$$N(\varepsilon) = \mathrm{O}\Bigg(\underbrace{C_n \inf_{\nu\in[0,1]}\left(\frac{L_\nu^2}{\mu^{1+\nu}\varepsilon^{1-\nu}}\right)^{1/(1+\nu)}}_{\substack{\text{число итераций}\\ \text{на одном рестарте}}}\underbrace{\left\lceil \ln\left(\frac{\mu\|x_*-x^0\|^2}{\varepsilon}\right)\right\rceil}_{\text{число рестартов}}\Bigg) \to$$

$$\to \mathrm{O}\Bigg(\underbrace{C_n \inf_{\nu\in[0,1]}\left(\frac{L_\nu^2}{\mu^{1+\nu}\varepsilon^{1-\nu}}\right)^{1/(1+3\nu)}}_{\substack{\text{число итераций}\\ \text{на одном рестарте}}}\underbrace{\left\lceil \ln\left(\frac{\mu\|x_*-x^0\|^2}{\varepsilon}\right)\right\rceil}_{\text{число рестартов}}\Bigg),$$

$$\delta = \mathrm{O}(\varepsilon) \to \widetilde{\mathrm{O}}\left(\frac{\varepsilon}{N(\varepsilon)}\right).$$

[1] Заметим, что у метода линейного каплинга (МЛК) [127] за счёт наличия двух проектирований на каждой итерации имеются некоторые дополнительные свойства (по сравнению с методом подобных треугольников, у которого одно проектирование), обнаруженные недавно [16, 267, 268, 322, 323, 470]. К сожалению, пока не удалось предложить такой вариант МЛК, который мог бы работать с моделью функции из § 3, но при этом обладал бы отмеченными выше дополнительными свойствами. Отметим также, что у МЛК в варианте работы [127] при $Q = \mathbb{R}^n$ Grad-шаг лучше заменить на Mirr-шаг, чтобы в случае неевклидовой прокс-структуры гарантировать ввиду формулы (3.16) равномерную ограниченность последовательности, генерируемой методом [4]. Другими словами, при $Q = \mathbb{R}^n$ соотношения (1.35), (1.36) стоит заменять на (2.29) с соответствующим выбором размеров шагов.

Данные оценки являются неулучшаемыми (оптимальными) [80, приложение А. С. Немировского], [233, 326, 462].

♦ В свою очередь, эти оценки можно обобщить на так называемые *промежуточные методы* [233, гл. 6], [271], которые представляют собой выпуклые комбинации неускоренного и ускоренного градиентного метода: в выписанных формулах вместо $\left[\frac{1}{1+\nu}; \frac{1}{1+3\nu}\right]$ следует писать $\frac{1}{1+\nu+2p\nu}$, при этом

$$\delta = \widetilde{\mathrm{O}}\Big(\frac{\varepsilon}{N(\varepsilon)^p}\Big), \quad p \in [0,1].$$

Такого рода обобщения могут потребоваться, например, при решении задач оптимизации в гильбертовых пространствах [292]. Детали и дальнейшее обобщение на случай неточного проектирования (см. формулу (3.3) и упражнение 3.7) имеются в работе [269]. ♦

Для большей наглядности далее (если не оговорено противное) рассматривается задача выпуклой безусловной оптимизации:

$$f(x) \to \min_{x \in \mathbb{R}^n}.$$

В качестве нормы выбирается 2-норма. В качестве прокс-функции берётся $d(x) = \frac{1}{2}\|x\|_2^2$, см. § 2.

♦ Тем не менее стоит отметить, что всё написанное далее с точностью до логарифмических множителей (см. константу C_n в упражнении 2.3 и в конце § 5) переносится и на задачи выпуклой оптимизации на множествах простой структуры. Исключением являются неполноградиентные методы для гладких задач выпуклой оптимизации. На данный момент для таких методов не удалось полностью перенести основные известные сейчас результаты для безусловных гладких задач на гладкие задачи оптимизации на множествах простой структуры [16, 20, 21, 267, 268, 487, 538].

Так же как и в основном тексте пособия, далее можно считать, что все константы, характеризующие оптимизируемую функцию, относятся не ко всему пространству, а только к шару с центром в точке старта и радиусом, равным (с точностью до логарифмического множителя) расстоянию от точки старта до (ближайшего) решения [20, 32, 260, 267, 268, 272, 423]. ♦

Изложенные выше результаты (в том числе и в ускоренном случае) с помощью *минибатчинга* (mini-batching) переносятся на задачи

стохастической оптимизации[2] [31, 51, 74, 86, 96, 98, 186, 233, 256, 271]. Заметим, что конструкция минибатчинга позволяет не только переносить оптимальные методы для детерминированных задач на задачи стохастической оптимизации с сохранением свойства оптимальности, но и получать оптимальные методы (по числу вызовов оракула) для задач стохастической оптимизации из неоптимальных методов для детерминированных задач. Опишем вкратце в простейшем случае суть конструкции. В задачах стохастической оптимизации вместо градиента функционала $\nabla f(x)$ оракул выдаёт его несмещённую оценку (*стохастический градиент*) $\nabla_x f(x, \xi)$ с конечной дисперсией D:

$$E_\xi[\nabla_x f(x, \xi)] \equiv \nabla f(x), \quad E_\xi\left[\|\nabla_x f(x, \xi) - \nabla f(x)\|_2^2\right] \leqslant D.$$

♦ Заметим, что приводимые далее результаты можно распространить на более общую концепцию стохастического шума [278, 309, 382, 516, 566, 578, 594] (обобщённое условие слабого и сильного роста):

$$E_\xi\left[\|\nabla_x f(x, \xi)\|_2^2\right] \leqslant D + c_1\|\nabla f(x)\|_2^2 + c_2 L \cdot (f(x) - f(x_*)).$$

Отметим, что в машинном обучении в последнее время достаточно популярны *перепараметризованные модели*, в которых выполняется *условие интерполяции*: $\nabla_x f(x_*, \xi) \cong 0$ тождественно по ξ [594]. В этом случае имеет место условие слабого роста ($D = 0$, $c_1 = 0$) с $c_2 = 2$. Некоторые частные случаи этой концепции будут далее разобраны (безградиентные методы). Отметим также, что в статистической теории обучения встречаются и другие показатели степени, в частности, замена

$$\|\nabla f(x)\|_2^2 \to \|\nabla f(x)\|_2^{1/2}$$

отвечает условию малого шума Цыбакова — Массара и Бернштейна; см. работу [154]. ♦

Конструкция *минибатчинга* заключается в подстановке в метод вместо неизвестного градиента $\nabla f(x)$ его (вычислимой) оценки

$$\overset{r}{\nabla}_x f(x, \{\xi^l\}_{l=1}^r) = \frac{1}{r}\sum_{l=1}^r \nabla_x f(x, \xi^l),$$

где $\{\xi^l\}_{l=1}^r$ — независимые одинаково распределённые (так же, как ξ) случайные величины, и правильном выборе параметра r. Выбрать

[2] Можно обойтись и без минибатчинга, также можно рассмотреть и седловые задачи (и даже вариационные неравенства), см., например, [32, 84, 149, 164, 186, 233, 258, 333, 348, 351, 578]. Отметим, что в работе [348] описывается метод, который может работать и в условиях отсутствия точных знаний о параметре D (см. также [149, 231, 351, 379, 426]).

этот параметр помогают следующие два неравенства (здесь $L = L_1$ в обозначениях (2.4)):

$$E_{\{\xi^l\}_{l=1}^r}\Big[\Big\|\overset{r}{\nabla}_x f(x,\{\xi^l\}_{l=1}^r) - \nabla f(x)\Big\|_2^2\Big] \leqslant \frac{D}{r},$$

$$\Big\langle \nabla f(x) - \overset{r}{\nabla}_x f(x,\{\xi^l\}_{l=1}^r), \mathrm{v}\Big\rangle \leqslant \underbrace{\frac{1}{2L}\Big\|\overset{r}{\nabla}_x f(x,\{\xi^l\}_{l=1}^r) - \nabla f(x)\Big\|_2^2}_{\delta} + \frac{L}{2}\|\mathrm{v}\|_2^2,$$

а также результаты о сходимости исследуемого метода при наличии неточного оракула (подобно § 2). Последнее неравенство можно переписать более удобным образом (см. неравенство (2.3)):

$$f(x^{k+1}) \leqslant f(x^k) + \Big\langle \overset{r}{\nabla}_x f\Big(x^k, \{\xi^{k+1,l}\}_{l=1}^r\Big), x^{k+1} - x^k\Big\rangle + \frac{2L}{2}\|x^{k+1} - x^k\|_2^2 + \delta^{k+1}.$$

Используя это неравенство в цепочке рассуждений (2.10)–(2.12), придём к следующему аналогу неравенства (2.12):

$$h\Big\langle \overset{r}{\nabla}_x f\Big(x^k, \{\xi^{k+1,l}\}_{l=1}^r\Big), x^k - x\Big\rangle \leqslant$$
$$\leqslant h\cdot\Big(f(x^k) - f(x^{k+1}) + \delta^{k+1}\Big) + \frac{1}{2}\|x - x^k\|_2^2 - \frac{1}{2}\|x - x^{k+1}\|_2^2,$$

где $h = 1/(2L)$. Взяв от обеих частей этого неравенства условное математическое ожидание $E_{x^{k+1}}[\cdot \mid x^1, \ldots, x^k]$, получим

$$f(x^k) - f(x) \leqslant \langle \nabla f(x^k), x^k - x\rangle \leqslant$$
$$\leqslant f(x^k) - E_{x^{k+1}}[f(x^{k+1}) \mid x^1, \ldots, x^k] + E_{x^{k+1}}[\delta^{k+1} \mid x^1, \ldots, x^k] +$$
$$+ L\|x - x^k\|_2^2 - E_{x^{k+1}}[L\|x - x^{k+1}\|_2^2 \mid x^1, \ldots, x^k].$$

Суммируя выписанные неравенства и вычисляя полное математическое ожидание, можно получить при $x = x_*$ аналог неравенства (2.22) с $L := 2L$. Исходя из неравенства (2.22) будем выбирать r следующим образом:

$$\frac{D}{2Lr} \simeq E[\delta] = \frac{\varepsilon}{2} \;\Rightarrow\; r \simeq \max\Big\{\frac{D}{L\varepsilon}, 1\Big\}.$$

Поскольку общее количество итераций (см. § 2 и теорему 3.1) равно $\mathrm{O}(LR^2/\varepsilon)$, общее число обращений к оракулу за стохастическим градиентом $\nabla_x f(x,\xi)$ при не малых значениях D будет равно

$$N(\varepsilon) = \mathrm{O}\Big(\frac{DR^2}{\varepsilon^2}\Big).$$

Эта же оценка получается и для ускоренных методов. Данная оценка является неулучшаемой оценкой для класса задач выпуклой стохастической оптимизации [74, 111, 554].

♦ В случае, если не известны значения L и(или) D, можно использовать следующий приём: с помощью подбора L^{k+1} (см. § 5) добиваемся выполнения неравенства

$$f(x^{k+1}) \leqslant f(x^k) + \left\langle \overset{r^{k+1}}{\nabla}_x f\left(x^k, \{\xi^{k+1,l}\}_{l=1}^{r^{k+1}}\right), x^{k+1} - x^k \right\rangle + \frac{2L^{k+1}}{2}\|x^{k+1} - x^k\|_2^2 + \frac{\varepsilon}{2}$$

с $r^{k+1} \simeq D_0/(L^{k+1}\varepsilon)$, где D_0 — оценка сверху для D ($D_0 \geqslant D$). Здесь также считаем, что $L \leqslant D_0/\varepsilon$. Описанный способ подбора L^{k+1} позволяет аналогично изложенному выше получить следующее неравенство:

$$\frac{1}{2L^{k+1}}\left\langle \overset{r^{k+1}}{\nabla}_x f\left(x^k, \{\xi^{k+1,l}\}_{l=1}^{r^{k+1}}\right), x^k - x \right\rangle \leqslant$$
$$\leqslant \frac{1}{2L^{k+1}} \cdot \left(f(x^k) - f(x^{k+1}) + \frac{\varepsilon}{2}\right) + \frac{1}{2}\|x - x^k\|_2^2 - \frac{1}{2}\|x - x^{k+1}\|_2^2.$$

Далее нужно специальным образом выбрать критерий останова метода. Метод останавливается, когда суммарное (по всем итерациям) число вычислений стохастического градиента достигнет некоторого заданного уровня $\widetilde{N}(\varepsilon) = \text{const} \cdot D_0 R^2/\varepsilon^2$. На практике так делать не обязательно. На последней итерации в общем случае уже не получится подобрать r^{k+1} по L^{k+1}, поскольку r^{k+1} будет определяться критерием останова, поэтому на последней итерации, наоборот, по r^{k+1} будет выбрано $L^{k+1} \simeq D_0/(r^{k+1}\varepsilon)$. В силу выпуклости функции $f(x)$ имеем

$$\sum_k \frac{1}{2L^{k+1}}\left\langle \overset{r^{k+1}}{\nabla}_x f\left(x^k, \{\xi^{k+1,l}\}_{l=1}^{r^{k+1}}\right), x^k - x \right\rangle \geqslant$$
$$\geqslant \sum_k \frac{1}{2L^{k+1}} \cdot (f(x^k) - f(x)) +$$
$$+ \sum_k \frac{1}{2L^{k+1}}\left\langle \overset{r^{k+1}}{\nabla}_x f\left(x^k, \{\xi^{k+1,l}\}_{l=1}^{r^{k+1}}\right) - \nabla f(x^k), x^k - x \right\rangle \simeq$$
$$\simeq \sum_k \frac{1}{2L^{k+1}} \cdot (f(x^k) - f(x)) +$$
$$+ \frac{\varepsilon}{2D_0} \sum_{i=0}^{\widetilde{N}(\varepsilon)} \left\langle \nabla_x f(x^{k(i)}, \xi^i) - \nabla f(x^{k(i)}), x^{k(i)} - x \right\rangle.$$

Если бы последняя сумма была суммой мартингал-разностей, что, в частности, означало бы, что её полное математическое ожидание равно нулю, то описанный выше способ выбора шага $h^{k+1} = 1/(2L^{k+1})$ и размера батча $r^{k+1} \simeq D_0/(L^{k+1}\varepsilon)$ гарантировал бы, что метод сойдётся за $O(LR^2/\varepsilon)$ итераций, см., например, [233, гл. 7], [260, 262, 271]. Хотя в численных экспериментах и наблюдалось хорошее соответствие

результатов данным оценкам (см. работы [259, 498], в которых также рассматривается ускоренный вариант описанной здесь процедуры), строго доказать всё это не удалось.

Заметим, что при практической реализации метода, когда мы увеличиваем $L^{k+1} := 2L^{k+1}$ (см. § 5), не следует генерировать новые стохастические градиенты в соответствующем (меньшем) количестве $r^{k+1} \simeq D_0/(L^{k+1}\varepsilon)$. В ходе процедуры $L^{k+1} := 2L^{k+1}$ можно использовать один и тот же изначально посчитанный на данной итерации вектор

$$\overset{r^{k+1}}{\nabla_x} f\left(x^k, \{\xi^{k+1,l}\}_{l=1}^{r^{k+1}}\right).$$

Отметим также, что в подавляющем большинстве приложений точные значения $f(x^k)$, $f(x^{k+1})$, как правило, недоступны, если недоступны соответствующие градиенты (про автоматическое дифференцирование будет сказано ниже). Доступны обычно случайные несмещённые реализации[3] значений целевой функции $f(x^k, \xi)$, $f(x^{k+1}, \xi)$. В этом случае (при дополнительном предположении о выпуклости функции $f(x, \xi)$ по x для всех ξ) вместо точных значений $f(x^k)$, $f(x^{k+1})$ в описанную выше адаптивную процедуру следует подставлять их оценки

$$f(x, \{\xi^l\}_{l=1}^r) = \frac{1}{r}\sum_{l=1}^r f(x, \xi^l), \quad x = x^k, x^{k+1},$$

которые построены аналогично оценке стохастического градиента. При этом теперь под L следует понимать другую константу, а именно не константу Липшица градиента $f(x) := E_\xi[f(x, \xi)]$, а худшую (по ξ) из констант Липшица градиентов (по x) функций $f(x, \xi)$. В этом случае можно гарантировать, что неравенства

$$f(x^{k+1}, \xi) \leqslant f(x^k, \xi) + \langle \nabla_x f(x^k, \xi), x^{k+1} - x^k\rangle + \frac{2L^{k+1}}{2}\|x^{k+1} - x^k\|_2^2 + \frac{\varepsilon}{2},$$

а следовательно, и

$$f\left(x^{k+1}, \{\xi^{k+1,l}\}_{l=1}^{r^{k+1}}\right) \leqslant f\left(x^k, \{\xi^{k+1,l}\}_{l=1}^{r^{k+1}}\right) + \\ + \left\langle \overset{r^{k+1}}{\nabla_x} f\left(x^k, \{\xi^{k+1,l}\}_{l=1}^{r^{k+1}}\right), x^{k+1} - x^k\right\rangle + \frac{2L^{k+1}}{2}\|x^{k+1} - x^k\|_2^2 + \frac{\varepsilon}{2}$$

заведомо выполняются, если L^{k+1} в процессе подбора дойдёт до значения этой новой константы L. Так же как и раньше, знать настоящее значение этой константы для реализации метода не обязательно.

[3] Часто говорят просто о реализациях. То, что они случайные и несмещённые, подразумевается.

К сожалению, построить законченную теорию здесь пока не получилось. Тем не менее особо отметим работу [149], в которой предлагается универсальный метод для решения монотонных стохастических вариационных неравенств на базе проксимального зеркального метода (см. замечание 5.1). По сути, используется стандартный проксимальный зеркальный метод, в котором L предлагается выбирать не так, как в замечании 5.1, а специальным образом, схожим со способом, использующимся в AdaGrad [230, 258]. Естественно, такой метод можно использовать и для решения седловых задач и задач выпуклой оптимизации [137, 276, 379]. С точностью до логарифмических множителей такой метод сходится (и по числу итераций, и по числу параллельных обращений к оракулу) для задач выпуклой оптимизации по оценкам, которые были выписаны выше для стохастического варианта градиентного спуска. Однако этот метод не является полностью адаптивным, поскольку, так же как и в AdaGrad, в стратегии выбора шага существенно используется информация о размере решения. Полностью адаптивный метод решения гладких стохастических монотонных вариационных неравенств был построен (с небольшими оговорками) в работе [351]. Данная работа представляется достаточно интересной в плане возможности перенесения полученных в ней результатов на ускоренные методы решения задач гладкой стохастической выпуклой оптимизации.

Отметим также работы [230, 600], в которой небольшая модификация метода AdaGrad исследуется с точки зрения глобальной сходимости к экстремуму для невыпуклых задач. ♦

Для μ-сильно выпуклой в 2-норме функции $f(x)$ приведённую оценку

$$N(\varepsilon) = \mathrm{O}\Big(\frac{DR^2}{\varepsilon^2}\Big)$$

можно улучшить с помощью рестартов (см. указание к упражнению 2.3 и конец § 5):

$$N(\varepsilon) = \mathrm{O}\Big(\min\Big\{\frac{DR^2}{\varepsilon^2}, \frac{D}{\mu\varepsilon}\Big\}\Big).$$

Данная оценка является неулучшаемой оценкой для класса задач сильно выпуклой стохастической оптимизации [74, 111]. Заметим, что не сильно выпуклую часть оценки можно получить из сильно выпуклой с помощью регуляризации $\mu \simeq \varepsilon/R^2$ (см. замечание 4.1). На более специальных классах задач приведённые оценки допускают уточнения. Например, сходимость может характеризоваться не дисперсией стохастического градиента (одно число), а его корреляционной матрицей [427, 454, 514, 534].

♦ Заметим, что для μ-сильно выпуклой в 2-норме функции $f(x)$ имеет место достаточно неожиданный результат о сходимости обычного градиентного спуска, в котором вместо градиента используется стохастический градиент (т. е. размер батча $r = 1$), с той же скоростью, что и обычный градиентный спуск (см. § 1, 2) в $\sqrt{D/(L\mu)}$-окрестность решения [398, 399, 454, 516]. На базе этого результата и стандартных приёмов (минибатчинга, каталиста, регуляризации) также можно получать приведённые выше оценки. ♦

Негладкий случай (см. (2.4) с $\nu = 0$) с помощью искусственного введения неточности в оракул (см. § 2) можно свести к гладкому случаю с $L \sim L_0^2/\varepsilon$. Поэтому (также оптимальную) оценку на число обращений к оракулу за стохастическим (суб-)градиентом в негладком случае можно записать в виде (следует сравнить с оценками из упражнений 2.1, 2.3)

$$N(\varepsilon) = \mathrm{O}\left(\min\left\{\frac{(L_0^2 + D)R^2}{\varepsilon^2}, \frac{L_0^2 + D}{\mu\varepsilon}\right\}\right).$$

Заметим, что *метод стохастического зеркального спуска* (2.19) (см. также упражнение 2.6) с $h = \varepsilon/M^2$, где $M^2 = L_0^2 + D$, и заменой $\nabla f(x^k) \to \nabla_x f(x^k, \xi^k)$ сходится в чезаровском смысле согласно первому аргументу минимума в приведённой оценке [366]. В адаптивном варианте шаг требуется выбирать более изощрённым образом, чем просто заменой градиента на стохастический градиент в формулах из упражнения 2.6. Адаптивный вариант получил название *AdaGrad* [230, 258] и уже упоминался нами ранее. Различные варианты этого метода [231, 426] являются сейчас одними из основных алгоритмов обучения глубоких нейронных сетей [40]. С помощью рестартов, подобно упражнению 2.3, можно получить вариант метода стохастического зеркального спуска для сильно выпуклых задач [366, 369]. Впрочем, для евклидовой прокс-структуры существуют и прямые (без рестартов) варианты метода зеркального спуска для сильно выпуклых задач, см., например, работу [333] и цитированную там литературу.

Резюмируем приведённые выше результаты в виде табл. 2. В таблице приведены оценки на число итераций $N(\varepsilon)$ (вызовов градиентного и стохастического градиентного оракула), необходимых для решения задачи (в среднем) с точностью ε по функции:

$$E[f(x^N)] - f(x_*) \leqslant \varepsilon.$$

Во всех этих оценках под R можно понимать евклидово расстояние от точки старта до ближайшего к точке старта решения. Эти оценки

Таблица 2

$N(\varepsilon)$	$f(x)$ выпуклая	$f(x)$ μ-сильно выпуклая
$E\left[\|\nabla_x f(x,\xi)\|_2^2\right] \leqslant M^2$	$\dfrac{M^2R^2}{\varepsilon^2}$	$\dfrac{M^2}{\mu\varepsilon}$
$\|\nabla f(y)-\nabla f(x)\|_2 \leqslant L\|y-x\|_2$	$\sqrt{\dfrac{LR^2}{\varepsilon}}$	$\sqrt{\dfrac{L}{\mu}}\left\lceil\ln\left(\dfrac{\mu R^2}{\varepsilon}\right)\right\rceil$
$\|\nabla f(y)-\nabla f(x)\|_2 \leqslant L\|y-x\|_2$ $E\left[\|\nabla_x f(x,\xi)-\nabla f(x)\|_2^2\right] \leqslant D$	$\max\left\{\sqrt{\dfrac{LR^2}{\varepsilon}}, \dfrac{DR^2}{\varepsilon^2}\right\}$	$\max\left\{\sqrt{\dfrac{L}{\mu}}\left\lceil\ln\left(\dfrac{\mu R^2}{\varepsilon}\right)\right\rceil, \dfrac{D}{\mu\varepsilon}\right\}$

являются нижними оценками (не могут быть улучшены) с одними числовыми множителями и достигаются на описанных выше методах (их ускоренных вариантах) с другими числовыми множителями.

В табл. 2 числовые множители опущены. Сильно выпуклая строчка таблицы получается из выпуклой с помощью рестартов (см. упражнение 2.3 и конец § 5), в обратную сторону переход осуществляется с помощью регуляризации, см. замечание 4.1. Негладкий столбец таблицы будет иметь такой же вид и в случае детерминированного оракула, выдающего (суб-)градиент. Под M следует в этом случае понимать константу Липшица функционала. Негладкий столбец получается из гладких столбцов с помощью конструкции универсального метода, см. § 5. Стохастический гладкий столбец получается из детерминированного гладкого с помощью минибатчинга, как описано в этом приложении. Ускоренные оценки последних двух столбцов (с точностью до логарифмического множителя, зависящего от желаемой точности) могут быть получены из неускоренных с помощью конструкции каталист, см. замечание 3.3. Результаты, приведённые в табл. 2, можно обобщить на случай других норм, при этом $R^2 = 2V(x_*, x^0)$ (если решение не единственно, в этой формуле выбирается то из решений x_*, которое минимизирует правую часть), а в последней строчке и последнем столбце в оценках скорости сходимости используемых методов появятся дополнительные логарифмические множители $\ln n$, см. конец § 2, указание к упражнению 2.3 и неравенство больших уклонений в неевклидовом случае, приведённое ниже. Насколько нам известно, вопрос о том, возможно ли убрать эти множители, остаётся открытым. Скорее всего, ответ тут будет отрицательным, т. е. эти множители в общем случае убрать нельзя.

Заметим, что для задач выпуклой (стохастической) оптимизации таблицу, аналогичную табл. 2, можно построить для неполноградиент-

ных методов [9, 16, 257, 267, 268, 272, 415, 551, 552]; в гладком случае для сходимости по норме градиента целевого функционала [198, 283, 388, 472] (или норме градиентного отображения для целевого функционала [122]); для методов параллельной оптимизации [188, 237, 261, 605]; для методов распределённой оптимизации [142, 171, 261, 262, 307, 308]; для стохастических вариационных неравенств [393].

♦ В работе [122] предложена оригинальная техника итеративной регуляризации (см. замечание 4.1, а также работу [225], в которой отмеченная техника проинтерпретирована с помощью неускоренного проксимального градиентного спуска) исходной задачи в сочетании с горячим стартом (warm start). Огрубляя детали для большей наглядности, опишем вкратце эту технику в несильно выпуклом случае. Регуляризуем задачу с помощью $\sigma_1\|x-x^0\|_2^2$, где $\sigma_1 \sim \widetilde{O}(\varepsilon^2)$. Решаем оптимальным методом (с оракулом, выдающим стохастический градиент) регуляризованную задачу (см. табл. 2) до момента, когда невязка по функции уменьшится в два раза. То, что получаем на выходе, обозначим через x^1. Затем действуем по индукции: регуляризуем (дополнительно) задачу с прошлой итерации с помощью $\sigma_{k+1}\|x-x^k\|_2^2$, где $\sigma_{k+1} = 2\sigma_k$, и решаем её, стартуя из точки x^k, с удвоенной точностью по функции. После $K = O(\log_2(C/\varepsilon^2))$ таких итераций (перезапусков) получим точку x^K, которая (с точностью до размерного множителя) будет $\widetilde{O}(\varepsilon^2)$-решением исходной задачи по функции (на самом деле в этом месте мы намеренно существенно огрубили описание конструкции работы [122], что, впрочем, практически не влияет на сам метод), а следовательно (см. формулу (1.8)), и $\widetilde{O}(\varepsilon)$-решением исходной задачи по критерию малости нормы градиента (градиентного отображения). Поскольку согласно табл. 2 и способу согласованного (пропорционального) увеличения параметра сильной выпуклости и уменьшения невязки по функции на каждой итерации требуется $\widetilde{O}(\varepsilon^{-2})$ раз вычислять стохастический градиент, общая трудоёмкость описанного подхода также будет равна $\widetilde{O}(\varepsilon^{-2})$. В данном случае несложно понять, что оценка оптимальная [122] (с точностью до логарифмических множителей). ♦

Насколько нам известно, на данный момент открытым остаётся вопрос об оптимальных оценках на число вызовов оракула (при оптимальных оценках на число коммуникаций) на один узел для задач распределённой стохастической децентрализованной оптимизации. Имеется гипотеза, дающая ответ на этот вопрос [261, 262].

Открытым остаётся вопрос и об оптимальных оценках для параллельных алгоритмов в централизованной оптимизации. В работе [605]

было показано, что если r узлов (машин) параллельно могут решать исходную задачу стохастической оптимизации, синхронизируясь (через центральный узел) не чаще, чем через каждые l итераций (вычислений стохастического градиента), то после T таких l-циклов для широкого класса задач гладкой выпуклой стохастической оптимизации и широкого класса допустимых алгоритмов имеет место следующая нижняя оценка на качество выдаваемого решения $\tilde{x}^T$ (числовые множители опущены для наглядности):

$$E[f(\tilde{x}^T)] - f(x_*) \geqslant \frac{LR^2}{(lT)^2} + \sqrt{\frac{DR^2}{rlT}},$$

Активные исследования ведутся в направлении разработки таких параллельных централизованных алгоритмов, которые как можно лучше приближаются к этой нижней границе [606]. На данный момент не известны методы, работающие в общем случае согласно приведённой нижней оценке (см. также нижнюю оценку в немного более общем случае [604]). Основная мотивация здесь — это приложения к Federated Learning [372]. Сюда можно отнести и набирающие популярность исследования по local SGD [378, 383, 390, 565, 604, 606]. Особо отметим работу [390], в которой построена общая теория сходимости неускоренного стохастического градиентного спуска (с консенсусом) при самых разнообразных предположениях об архитектуре (параллельная, типа local SGD, gossip, децентрализованная распределённая, меняющаяся со временем и т. д.).

Замечание 1. Во многих задачах (в частности, в задачах анализа данных [40, 547, 608]) функционал имеет вид суммы большого числа слагаемых:

$$f(x) = \frac{1}{m}\sum_{l=1}^{m} f_l(x) \to \min_{x \in Q}.$$

Если m — очень большое число, то вместо честного и дорогого вычисления градиента вычисляют стохастический градиент, случайно (равновероятно) выбирая $r \ll m$ слагаемых $\{\xi^l\}_{l=1}^r$ и формируя стохастический градиент (несмещённую оценку градиента) по формуле

$$\nabla f(x, \{\xi^l\}_{l=1}^r) = \frac{1}{r}\sum_{l=1}^{r} \nabla f_{\xi^l}(x).$$

Такой подход называют *методом рандомизации суммы*, см., например, [20]. С другой стороны, описанную конструкцию можно понимать и как минибатчинг, если посмотреть на исходную постановку

задачи следующим образом:

$$f(x) = E_\xi[f(x,\xi)] \to \min_{x\in Q}, \quad f(x,\xi) = f_\xi(x),$$
$$\nabla_x f(x,\xi) = \nabla f_\xi(x), \quad P(\xi = l) = \frac{1}{m}, \quad l = 1,\dots,m.$$

Именно в таком ключе обычно смотрят на минибатчинг в глубоком обучении [40, 533]. На практике иногда можно получить дополнительное ускорение, выбирая случайно $\{\xi^l\}_{l=1}^r$ без повторений [115, 449, 518].

Отметим, что минибатчинг хорошо параллелится, в отличие от процедур типа стохастического усреднения (stochastic averaging) [44]. В популярной работе [494] обсуждается альтернативная возможность распараллеливания стохастического градиентного спуска, в которой отсутствует синхронизационная накладка и обсуждается возможность одновременной записи в общую память. Интересные исследования по альтерантивным к минибатчингу архитектурам параллелизации (например, local SGD) описаны в работах [302, 378, 383, 390, 565, 604–606] и цитированной в них литературе.

Продемонстрируем, насколько важную роль играет стохастическая оптимизация (рандомизация) при решении задач «больших размеров» следующим примером.

Вернёмся к приведённой выше задаче минимизации суммы. Ограничимся рассмотрением выпуклого (но не сильно выпуклого) случая. Воспользуемся методом рандомизации суммы с $r = 1$. Пусть все функции $f_l(x)$ в определении функционала $f(x)$ имеют ограниченные константы L_0 (см. формулу (2.4)). Тогда для решения исходной задачи минимизации суммы с точностью по функции (в среднем) ε требуется $O(L_0^2R^2/\varepsilon^2)$ обращений к оракулу за (суб-)градиентами случайно выбранных слагаемых $f_l(x)$. Здесь R — расстояние в 2-норме от точки старта до (ближайшего к точке старта) решения. В то же время, даже если все слагаемые достаточно гладкие — имеют ограниченные константы L_1 (см. формулу (2.4)), для достижения той же точности ε быстрому градиентному методу потребуется $O(m\sqrt{L_1R^2/\varepsilon})$ обращений к оракулу за градиентами[4] $f_l(x)$. Для задач с большим

[4] Заметим, что из условия $L_1 < \infty$ не следует, что $L_0 < \infty$ (можно рассмотреть пример квадратичной функции), и потому нет гарантий, что стохастический градиент (градиент случайно выбранного слагаемого) имеет конечный второй момент (дисперсию). Это создает определённые трудности, которые в общем случае являются препятствием для прямого применения (без использования приёма редукции дисперсии) ускоренных методов стохастической оптимизации для минимизации суммы функций [146], см. также п. 3 упражнения 1.8.

числом слагаемых при невысоких требованиях к точности решения первый (рандомизированный) способ может оказаться предпочтительнее. Отметим, что выписанная оценка $O(m\sqrt{L_1R^2/\varepsilon})$ ввиду наличия специальной структуры у задачи уже не будет оптимальной. Оптимальна следующая оценка, достигаемая на методах редукции дисперсии [109, 120, 124, 139, 403, 413, 433, 434, 607]: $O(m+\sqrt{mL_1R^2/\varepsilon})$. Невыпуклый случай с критерием малости нормы градиента рассматривается в [122, 123, 125, 126, 128, 277, 403, 412]). Оптимальные оценки были получены в [277], см. также [138, 509, 598, 599] и самый конец приложения. В случае, если дополнительно известно, что функция $f(x)$ является μ-сильно выпуклой в 2-норме, оптимальной оценкой будет $\widetilde{O}(m+\sqrt{mL_1/\mu})$; см. [406]. Наиболее современный обзор текущих достижений в данной области см. в [311, 403, 437, 618]. Приведённые выше результаты обобщаются на случай, когда оракул вместо градиента функции $f_l(x)$ выдаёт её стохастический градиент [21, 398, 399, 414]. Приведённые выше результаты также обобщаются на седловые задачи и вариационные неравенства [503, 617]. Перенесение на модельную общность на данный момент осуществлено лишь частично: для композитных и проксимальных случаев [228, 414].

♦ Строго говоря, следует различать константу L_1, входящую в оценку $O(m\sqrt{L_1R^2/\varepsilon})$, и константу L_1, входящую в оценку $O(m+\sqrt{mL_1R^2/\varepsilon})$. В первом случае под L_1 следует понимать константу Липшица градиента оптимизируемой функции $f(x)$. Во втором случае под L_1 следует понимать максимальную из констант Липшица градиентов слагаемых $f_l(x)$, $l=1,\ldots,m$. Во втором случае L_1 может быть заметно больше. Например, для функции ($m=n$)

$$f(x)=\frac{1}{n}\|x\|_2^2=\frac{1}{n}\sum_{l=1}^{n}x_l^2=\frac{1}{m}\sum_{l=1}^{m}f_l(x)$$

разница будет в $m=n$ раз [340, 576]. Впрочем, если в методах редукции дисперсии выбирать слагаемые в сумме не равновероятно (случайно), а со специальными вероятностями, зависящими от констант слагаемых липшицевых градиентов, то можно заменить в итоговой оценке сложности худшую константу слагаемых липшицевых градиентов на некоторую среднюю величину [328, 453]. ♦

Приведённые оценки достижимы лишь в случае, если имеется доступ к градиенту каждого слагаемого в отдельности [141], а не только к целому градиенту оптимизируемой функции, т. е. используется *инкрементальный метод* [109]. Далее в примере будет продемонстриро-

ван способ получения такого типа оценок. Тем не менее даже с учётом этого замечания можно привести конкретные примеры, когда первый способ по-прежнему остаётся предпочтительнее. Отметим также, что для конкретных примеров приведённые оценки (и их обобщения на негладкие выпуклые задачи [120]) могут быть улучшены [214].

Ещё раз подчеркнём, что при решении гладких выпуклых задач стохастической оптимизации или при решении гладких выпуклых задач рандомизированными методами с помощью конструкции минибатчинга удаётся эффективно распараллеливать вычисления. Скажем, в разобранном выше примере приведённая ранее оценка $O(L_0^2R^2/\varepsilon^2)$ на число вычислений градиентов случайно выбранных слагаемых $f_l(x)$ (при дополнительном предположении о гладкости функционала) может быть редуцирована (за счёт распараллеливания при минибатчинге) до оценки $O(\sqrt{L_1R^2/\varepsilon})$. В данном случае результат вполне ожидаем и достижим без минибатчинга, просто за счёт структуры функционала и возможности параллельно (на m процессорах) вычислять градиент в соответствующем быстром градиентном методе. Однако конструкция минибатчинга естественным образом распараллеливается во всех случаях, независимо от наличия у задачи дополнительной структуры. Здесь лишь хотелось продемонстрировать эффективность такого распараллеливания на простом примере.

Подробнее о возможности параллельного и распределённого решения задач выпуклой оптимизации вида суммы (в том числе, приходящих из анализа данных) можно посмотреть в работе [261].

Заметим, что, хотя методы редукции дисперсии дают в теории лучший результат в широком диапазоне параметров, на практике на задачах машинного обучения они могут проигрывать обычному методу стохастического градиентного спуска со специально выбранным размером шага и небольшим размером батча [594], особенно для невыпуклых задач [555]. В (сильно) выпуклом случае этому есть простое объяснение (см., например, [390] и цитированную там литературу), в основе которого лежит результат о том, что в оценку скорости сходимости стохастического градиентного спуска (Stochastic Gradient Descent (SGD)) для задачи вида суммы вместо дисперсии входит агрегат, который уменьшается (вплоть до нуля) с увеличением степени схожести слагаемых в сумме друг с другом (например, если выполняется упоминаемое выше условие интерполяции, приводящее к условию слабого роста). Последнее обстоятельство в разной степени присуще многим задачам машинного обучения. Другое объяснение связано с дополнительными возможностями для SGD в вы-

боре больших шагов и с меньшей чувствительностью к неизвестным параметрам задачи, таким как, например, константа сильной выпуклости [454]. ■

♦ Приведём, следуя работам [21], [186, п. 6.3], [416], для задачи минимизации суммы из замечания 1 с $Q = \mathbb{R}^n$ (а значит, $\nabla f(x_*) = 0$) схему способа (SVRG + каталист) получения оптимальных (с точностью до логарифмического множителя) оценок на число вычислений градиентов слагаемых. Будем считать, что все функции $f_k(x)$ имеют L_1-липшицев градиент и μ-сильно выпуклые (всё относительно 2-нормы). В качестве несмещённой оценки градиента возьмём вектор (*variance reduction* — выделение главной части в методах Монте-Карло [50, п. 3.1.1])

$$\nabla_x f(x^{s,k}, \xi^{s,k}) = \nabla f_{\xi^{s,k}}(x^{s,k}) - \nabla f_{\xi^{s,k}}(y^s) + \nabla f(y^s),$$

где

$$y^s = x_{\overline{N}}^{s-1} = \frac{1}{\overline{N}} \sum_{k=0}^{\overline{N}-1} x^{s-1,k},$$

а случайная величина ξ принимает равновероятно одно из значений $1, \ldots, m$, и будем использовать специальным образом рестартованный градиентный спуск с минибатчингом (см. выше)

$$x^{s,k+1} = x^{s,k} - \frac{1}{L_1} \frac{r^s}{\overline{N}} \nabla_x f(x^{s,k}, \xi^{s,k}),$$
$$k = 0, \ldots, \overline{N} - 1, \quad r^s \simeq \max\left\{\frac{D^s}{L_1 \varepsilon}, 1\right\},$$

где параметр $\overline{N}$ будет выбран позже как $\overline{N} = \mathrm{O}(4L_1/\mu)$,

$$E_\xi\left[\left\|\nabla_x f(x^{s,k}, \xi) - E_\xi[\nabla_x f(x^{s,k}, \xi)]\right\|_2^2\right] \leqslant E_\xi\left[\|\nabla_x f(x^{s,k}, \xi)\|_2^2\right] \leqslant$$
$$\leqslant \underbrace{2E_\xi\left[\left\|\nabla f_{\xi^{s,k}}(x^{s,k}) - \nabla f_{\xi^{s,k}}(x_*)\right\|_2^2\right]}_{\overset{(1.8)}{\leqslant} 2E_\xi[2L_1(f_{\xi^{s,k}}(x^{s,k}) - f_{\xi^{s,k}}(x_*) - \langle \nabla f_{\xi^{s,k}}(x_*),\, x^{s,k} - x_*\rangle)]} \overset{\nabla f(x_*)=0}{=} \underset{= 4L_1(f(x^{s,k}) - f(x_*))}{} +$$
$$+ \underbrace{2E_\xi\left\|\nabla f_{\xi^{s,k}}(y^s) - \nabla f_{\xi^{s,k}}(x_*) - \nabla f(y^s)\right\|_2^2}_{\overset{\nabla f(x_*)=0}{=} 2E_\xi\|\nabla f_{\xi^{s,k}}(y^s) - \nabla f_{\xi^{s,k}}(x_*) - E_\xi[\nabla f_{\xi^{s,k}}(y^s) - \nabla f_{\xi^{s,k}}(x_*)]\|_2^2 \leqslant} \leqslant$$
$$\underset{\leqslant 2E_\xi\|\nabla f_{\xi^{s,k}}(y^s) - \nabla f_{\xi^{s,k}}(x_*)\|_2^2 \overset{(1.8)}{\leqslant} 4L_1(f(y^s) - f(x_*))}{}$$
$$\leqslant 4L_1 \cdot (f(x^{s,k}) - f(x_*)) + 4L_1 \cdot (f(y^s) - f(x_*)) = \mathrm{O}(L_1 \Delta f^s) \overset{\text{def}}{=} D^s, \quad (*)$$

где $\Delta f^s = f(y^s) - f(x_*)$. Здесь по k идёт внутренний цикл, а по s — внешний[5].

Возьмём $\overline{N} = \mathrm{O}(4L_1/\mu)$, тогда

$$\Delta f^{s+1} = \max\left\{\mathrm{O}\left(\Delta f^s \exp\left(-\frac{\mu}{L_1}\overline{N}\right)\right),\ \mathrm{O}\left(\frac{D^s}{\mu\overline{N}}\right)\right\},$$

$$\frac{D^s}{\mu\overline{N}} = \mathrm{O}\left(\frac{L_1\Delta f^s}{\mu\overline{N}}\right) = \mathrm{O}\left(\frac{1}{4}\Delta f^s\right), \quad \Delta f^s \exp\left(-\frac{\mu}{L_1}\overline{N}\right) = \mathrm{O}\left(\frac{1}{4}\Delta f^s\right).$$

Получим (все рассуждения здесь и далее можно провести более аккуратно, убрав O( ))

$$\Delta f^{s+1} \leqslant \mathrm{O}\left(\frac{1}{2}\Delta f^s\right).$$

Таким образом, общее число вычислений $\nabla f_l(x)$, необходимых для достижения точности ε по функции, будет составлять

$$\mathrm{O}\left(\left(m + \frac{L_1}{\mu}\right)\cdot \ln\left(\frac{\Delta f^0}{\varepsilon}\right)\right).$$

Собственно, каталист и был впервые предложен в работе [433] как способ ускорения этой оценки до

$$\widetilde{\mathrm{O}}\left(\left(m + \sqrt{m\frac{L_1}{\mu}}\right)\cdot \ln\left(\frac{\Delta f^0}{\varepsilon}\right)\right).$$

Напомним, что в каталисте (см. замечание 3.3) идёт игра на выборе параметра регуляризации исходной задачи $L \gg \mu$. Сложность решения (число вычислений $\nabla f_l(x)$) на каждой (внешней) итерации каталиста внутренней задачи с необходимой точностью будет составлять

$$\mathrm{O}\left(\left(m + \frac{L_1}{\mu + L}\right)\cdot \ln\left(\frac{\Delta f^0}{\varepsilon}\right)\right) = \widetilde{\mathrm{O}}\left(m + \frac{L_1}{L}\right),$$

а число внешних итераций каталиста будет составлять (чтобы получить такую оценку, метод Монтейро — Свайтера из замечания 3.3 необходимо ещё рестартовать, см. конец § 5)

$$\mathrm{O}\left(\sqrt{\frac{L}{\mu}}\ln\left(\frac{\Delta f^0}{\varepsilon}\right)\right) = \widetilde{\mathrm{O}}\left(\sqrt{\frac{L}{\mu}}\right).$$

Выбирая

$$L = \mathrm{O}\left(\max\left\{\frac{L_1}{m}, \mu\right\}\right),$$

получим желаемый результат.

[5] С помощью соотношения $(*)$ и минибатчинга (см. начало приложения) можно получить обобщение описываемого здесь метода редукции дисперсии на случай, когда вместо градиента слагаемого доступен только его стохастический градиент. Ускоряя такой метод каталистом [399], можно получить оптимальные оценки сложности [398, 413].

Идея каталиста также будет продемонстрирована ниже на более простом примере. В качестве хорошего упражнения на использование техники каталист можно попробовать ускорить результаты работ [309, 329]. При этом результаты работы [309] интересно также было бы попробовать ускорить исходя из подхода, описанного в замечании 1.5 и тексте после него, см. также [578, 579]. ♦

Всё отмеченное выше переносится также на покомпонентные и безградиентные постановки задач[6] [9, 16, 20, 21, 32, 74, 86, 98, 117, 209, 268, 272, 359, 415, 469, 487, 537, 551]. Пусть случайный вектор e распределён, например, равномерно на евклидовой сфере в $\mathbb{R}^n$ радиуса 1, т. е. $\|e\|_2 = 1$, или равновероятно среди единичных ортов [39, 51, 561] (*покомпонентная рандомизация*). Тогда

$$\frac{f(x+\tau e)-f(x)}{\tau} \simeq \langle \nabla f(x), e\rangle, \quad E_e \underbrace{\big[n\langle \nabla f(x), e\rangle e\big]}_{\substack{\text{то, что подставляется}\\ \text{в метод вместо градиента}}} = \nabla f(x),$$

$$\langle \nabla f(x), \langle \nabla f(x), e\rangle e\rangle = \|\langle \nabla f(x), e\rangle e\|_2^2,$$

$$E_e\big[\|n\langle \nabla f(x), e\rangle e\|_2^2\big] = n\|\nabla f(x)\|_2^2.$$

Выписанные соотношения вкупе с основным неравенством (3.1) позволяют показать, что число итераций (число обращений к оракулу за значением функции или производной по направлению) для таких методов в среднем по порядку будет в n раз больше, чем для полноградиентных аналогов. В общем случае этот результат не может быть улучшен [74, 110]. Впрочем, при дополнительных предположениях улучшения возможны [9, 16, 21, 268, 488], см. также разобранный ниже пример решения задачи квадратичной оптимизации покомпонентным методом. Описанный результат вполне понятен, поскольку, запросив частные производные (или значения функции) по n координатным ортам, можно просто восстановить полный градиент. В этой связи отметим, что если есть доступ к программе (точнее, следует говорить о доступе к тексту программы, см. ниже), вычисляющей значение функции (так бывает далеко не во всех приложениях, см., например, [218]), то, как правило, лучше попробовать использовать *автоматическое дифференцирование* [495, гл. 8], чем просто аппроксимировать градиент (и тем более старшие производные) конечными разностями [46].

[6] Отметим, что про нижние оценки на скорость накопления малых неточностей неслучайной природы для неполноградиентных методов [526] на данный момент известно меньше, чем для полноградиентных методов [233].

♦ Автоматическое дифференцирование (automatic differentiation) — способ по программе, вычисляющей значение функции (дереву вычислений), построить программу, вычисляющую градиент функции и работающую не более чем в 4 раза дольше исходной. Однако такой способ требует в общем случае большей памяти — необходимо хранить в памяти всю историю (дерево) вычисления функции. Изначально результаты такого рода были получены (Баур — Штрассен) для полиномов, см. книгу [89] и цитированную там литературу. Впоследствии, в начале 80-х годов XX века, две группы в ЦЭМИ РАН [60] и ВЦ РАН (более полный и точный исторический обзор см. в книге [49] и цитированной там литературе) смогли получить описанный выше результат в наибольшей общности. Подробный обзор имеется также в работе [159]. В работе [473] приведён интересный пример использования автоматического дифференцирования для решения вспомогательных подзадач для методов 3-го порядка (см. ниже) с той же по порядку сложностью, что и для методов 2-го порядка (в частности, метода Ньютона), см. также указание к упражнению 3.10. Заметим, что аналогом автоматического дифференцирования для негладких выпуклых функций является лексикографическое дифференцирование, предложенное в конце 80-х годов XX века Ю. Е. Нестеровым [80, 478]. Интересно также заметить, что в ряде классических работ по нейронным сетям, в которых используется частный случай автоматического дифференцирования — *метод обратного распространения* (back propagation), имеются неточности. Эти неточности связанны как раз с тем, что для негладких функций (негладкость получается за счёт использования персептронов / функций активаций ReLu) используется процедура автоматического дифференцирования, обоснованно работающая только для гладких функций. ♦

Далее с помощью техники рестартов, следуя работе [21], мы объясним более точно, откуда в оценке скорости сходимости появляется множитель $\sim n$. Ключевое наблюдение базируется на формуле (1.8):

$$M^2 \simeq E_e\left[\|n\langle\nabla f(x), e\rangle e\|_2^2\right] = n\|\nabla f(x)\|_2^2 \leqslant 2Ln\cdot(f(x)-f(x_*))$$

и приведённой выше оценке скорости сходимости градиентного спуска с минибатчингом для задач стохастической оптимизации, см. также табл. 2 (здесь, как и раньше, мы работаем с точностью до поправки на вероятности больших уклонений или оговорки о том, что сходимость понимается в среднем, см., например, [96, 320]):

$$f(x^N)-f(x_*) = \mathrm{O}\left(\sqrt{\frac{M^2R^2}{N}}\right) \leqslant \mathrm{O}\left(\sqrt{\frac{2Ln\cdot(f(x^0)-f(x_*))R^2}{N}}\right)$$

Рестартуя метод каждый раз, когда происходит гарантированное (выписанной формулой) уполовинивание невязки по функции, получим следующую формулу для общего числа обращений к оракулу за значениями $\langle \nabla f(x), e\rangle$ или $\{f(x), f(x+\tau e)\}$ (см. также указания к упражнениям 1.3, 2.3):

$$N(\varepsilon) = \mathrm{O}\Big(n\frac{8LR^2}{\Delta f}\Big) + \mathrm{O}\Big(n\frac{8LR^2}{\Delta f/2}\Big) + \\ + \mathrm{O}\Big(n\frac{8LR^2}{\Delta f/4}\Big) + \ldots + \mathrm{O}\Big(n\frac{8LR^2}{\varepsilon}\Big) = n\cdot\mathrm{O}\Big(\frac{LR^2}{\varepsilon}\Big),$$

где $\Delta f = f(x^0) - f(x_*)$. К сожалению, по той же причине — ввиду грубости формулы (1.8), из-за которой не следует использовать для ускоренных методов рестарты по норме градиента для решения гладких сильно выпуклых задач (см. замечание 5.3), — здесь не получается похожим образом перенести описанную конструкцию на ускоренные градиентные спуски (см. упражнения 1.3, 5.7). Для ускоренных методов требуются более тонкие рассуждения [21, 272, 469, 487, 488].

Приведённые выше рассуждения можно повторить в случае μ-сильно выпуклой в 2-норме функции $f(x)$. При этом всё получится ещё проще. Можно не повторять рассуждения, а улучшить приведённую оценку с помощью рестартов (см. указание к упражнению 2.3 и конец § 5):

$$N(\varepsilon) = n\cdot\mathrm{O}\Big(\frac{L}{\mu}\Big\lceil\ln\Big(\frac{\mu R^2}{\varepsilon}\Big)\Big\rceil\Big).$$

Проверить соответствие данной оценки, аналогичной оценке, полученной ранее в несильно выпуклом случае, можно с помощью регуляризации $\mu \simeq \varepsilon/R^2$ (см. замечание 4.1).

До недавнего времени считалось, что так же просто, как это было описано выше в неускоренном случае, не удастся перенести описанные выше конструкции на ускоренные методы. Однако недавно было обнаружено [254, 433, 434] (см. также § 3), как приведённые выше оценки (и многие другие оценки скорости сходимости для неускоренных методов) могут быть единообразно перенесены на ускоренный случай с помощью новой довольно общей и вместе с тем достаточно простой техники *каталист*, см. замечание 3.3:

$$N(\varepsilon) = n\cdot\mathrm{O}\Bigg(\min\Bigg\{\sqrt{\frac{LR^2}{\varepsilon}}, \sqrt{\frac{L}{\mu}}\Big\lceil\ln\Big(\frac{\mu R^2}{\varepsilon}\Big)\Big\rceil\Bigg\}\Bigg).$$

Основным «структурным блоком» по-прежнему является вспомогательная задача (3.21),

$$x^{k+1} = \arg\min_{x\in\mathbb{R}^n}\Big\{f(x) + \frac{\widetilde{L}}{2}\|x - x^k\|_2^2\Big\},$$

которую необходимо решать на каждой итерации. Сложность решения этой задачи (число обращений к оракулу за значением функции/производной по направлению на одной итерации) неускоренным безградиентным/покомпонентным методом (с точностью до логарифмического множителя, см. упражнение 3.1) равна

$$\widetilde{\mathrm{O}}\left(n \cdot \frac{L+\widetilde{L}}{\mu+\widetilde{L}}\right).$$

С другой стороны, число внешних итераций ускоренного прокс-метода [97, 433, 434] с точностью до логарифмического множителя равно $\widetilde{\mathrm{O}}(\sqrt{\widetilde{L}/\mu})$. Для того чтобы получить такую оценку, метод Монтейро — Свайтера из замечания 3.3 необходимо ещё рестартовать, см. конец § 5. Оценка общего числа обращений к оракулу будет наилучшей и составит $\widetilde{\mathrm{O}}(n\sqrt{L/\mu})$ при выборе $\widetilde{L} \simeq L$, что соответствует в сильно выпуклом случае ранее приведённому результату.

Разобранный пример наглядно демонстрирует общую идею подхода: за счёт выбора параметра регуляризации $\widetilde{L}$ добиваться близкой к единице обусловленности вспомогательной задачи, тогда неоптимальность используемого для её решения подхода (напомним, что как раз для внутренней задачи используется неускоренный метод, который мы хотим ускорить) становится несущественной и за счёт ускоренности внешнего метода происходит общее ускорение рассматриваемой процедуры. Примечательно, что внешний ускоренный прокс-метод при данном подходе остаётся одним и тем же, в то время как внутренний неускоренный метод (для решения вспомогательной задачи) можно как угодно менять в зависимости от контекста. Отметим также, что выше рассуждения проводились с точностью до логарифмических множителей. К сожалению, если честно их выписывать, то выяснится, что такой подход приводит не к оптимальным оценкам, а к оценкам, оптимальным с точностью до логарифмических (по желаемой точности) множителей.

Хотя каталист и является универсальным способом ускорения всевозможных неускоренных методов, тем не менее на практике предпочитают использовать прямые ускоренные рандомизированные процедуры, см., например, [124, 413, 414] и замечание 2 ниже. Особенно активно в этом направлении работали З. Аллен-Зу [128] и Дж. Лан [410].

Важно также отметить, что даже в неускоренном случае приведённые выше рассуждения при покомпонентной рандомизации оказываются достаточно грубыми, поскольку не учитывают, что константу L теперь можно считать не по худшему направлению, а брать «средней» по всем направлениям, что может быть в $\sim \sqrt{n}$ раз меньше [21, 488].

Таблица 3

Метод линейного каплинга (МЛК)	Покомпонентный вариант МЛК (ПМЛК)
См. указание к упражнению 1.3 и замечание 1.6, $x^{k+1} = \tau z^k + (1-\tau)y^k,$ $y^{k+1} = x^{k+1} - \frac{1}{L}\nabla f(x^{k+1}),$ $z^{k+1} = z^k - h\nabla f(x^{k+1})$	$x^{k+1} = \tau z^k + (1-\tau)y^k$ Случайно и независимо разыгрываем $i_{k+1} \in [1,\dots,n]$ по правилу $P(i_{k+1}=i) = p_i \overset{\text{def}}{=} \frac{L_i^\beta}{\sum_{j=1}^n L_j^\beta},\ i=1,\dots,n,$ $y^{k+1}_{i_{k+1}} = x^{k+1}_{i_{k+1}} - \frac{1}{L_{i_{k+1}}}\frac{\partial f(x^{k+1})}{\partial x^{k+1}_{i_{k+1}}},$ $z^{k+1}_{i_{k+1}} = z^k_{i_{k+1}} - \frac{h}{p_{i_{k+1}}}\frac{\partial f(x^{k+1})}{\partial x^{k+1}_{i_{k+1}}}$

Замечание 2. Предположим, что для всех $x \in \mathbb{R}^n$, $i=1,\dots,n$, и $h \in \mathbb{R}$ выполнено неравенство

$$\left|\frac{\partial f(x+he_i)}{\partial x_i} - \frac{\partial f(x)}{\partial x_i}\right| \leqslant L_i h.$$

Пусть $\beta \in [0,1]$. Введём

$$\|x\|^2 = \sum_{i=1}^n L_i^{1-2\beta} x_i^2, \quad \breve{R}^2 = \frac{1}{2}\|x_* - x^0\|^2,$$

$$\|\nabla f(x)\|_*^2 = \sum_{i=1}^n L_i^{2\beta-1} \cdot \Big(\frac{\partial f(x)}{\partial x_i}\Big)^2.$$

Для МЛК согласно указанию к упражнению 1.3 имеем

$$\langle \nabla f(x^{k+1}), z^k - x_*\rangle \leqslant \frac{1}{2h}\|z^k - x_*\|_2^2 - \frac{1}{2h}\|z^{k+1} - x_*\|_2^2 + \frac{h\|\nabla f(x^{k+1})\|_2^2}{2},$$

т. е.

$$\langle \nabla f(x^{k+1}), z^k - x_*\rangle \leqslant \frac{1}{2h}\|z^k - x_*\|_2^2 - \frac{1}{2h}\|z^{k+1} - x_*\|_2^2 + Lh\cdot(f(x^{k+1}) - f(y^{k+1})).$$

Для ПМЛК аналогом приведённых неравенств будут неравенства

$$\frac{1}{p_{i_{k+1}}}\langle\langle \nabla f(x^{k+1}), e_{i_{k+1}}\rangle e_{i_{k+1}}, z^k - x_*\rangle \leqslant$$

$$\leqslant \frac{1}{2h}\|z^k - x_*\|^2 - \frac{1}{2h}\|z^{k+1} - x_*\|^2 + \frac{h\big\|\langle \nabla f(x^{k+1}), e_{i_{k+1}}\rangle e_{i_{k+1}}\big\|_*^2}{2p_{i_{k+1}}^2},$$

$$\langle \nabla f(x^{k+1}), z^k - x_* \rangle \leqslant$$
$$\leqslant \frac{1}{2h}\|z^k - x_*\|^2 - E_{i_{k+1}}\Big[\frac{1}{2h}\|z^{k+1} - x_*\|^2 \mid i_0, \ldots, i_k\Big] +$$
$$+ \check{L}h \cdot \big(f(x^{k+1}) - E_{i_{k+1}}[f(y^{k+1}) \mid |i_0, \ldots, i_k]\big),$$

где

$$\check{L} = \Big(\sum_{j=1}^{n} L_j^{\beta}\Big)^2.$$

Второе неравенство получается из первого взятием условного математического ожидания от обеих частей $E_{i_{k+1}}[\cdot \mid i_0, \ldots, i_k]$. Поскольку согласно указанию к упражнению 1.3 оценка числа итераций (вычислений градиента $f(x)$), необходимых для достижения по функции точности ε, имеет вид $O\big(\sqrt{LR^2/\varepsilon}\big)$, для ПМЛК естественно было бы ожидать, что оценка числа итераций (вычислений частных производных $f(x)$), необходимых для достижения по функции (в среднем) точности ε, будет составлять $O\big(\sqrt{\check{L}\check{R}^2/\varepsilon}\big)$. Так оно в действительности и оказывается [21, 118, 469, 488]. Аналогичные рассуждения можно было провести, взяв за основу метод подобных треугольников вместо МЛК [272].

Методы МЛК и ПМЛК можно осуществлять (с такими же оценками скорости сходимости) и без рестартов [21, 118, 127]. Для этого нужно выбирать (см. также замечание 1.6)

$$\tau_k = \frac{2}{k+2} \quad \text{и} \quad h_k = \frac{k+2}{2L} \quad \text{(МЛК)},$$
$$\tau_k = \frac{2}{k+2} \quad \text{и} \quad h_k = \frac{k+2}{2\check{L}} \quad \text{(ПМЛК)}.$$

Описанный покомпонентный метод имеет естественное блочно-покомпонентное обобщение [469]. При этом допускается, что рассматриваемую задачу оптимизации необходимо решать на множестве простой структуры, имеющем вид прямого произведения множеств, отвечающих различным блокам [21, 272]. ■

Поясним абзац, написанный непосредственно перед замечанием 2, простым примером, в котором сравним время работы быстрого градиентного метода, например МЛК из указания к упражнению 1.3, и его покомпонентного варианта — ПМЛК с $\beta = 1/2$ [21]. То же самое можно было продемонстрировать и для неускоренных методов. Итак, рассматривается задача квадратичной выпуклой оптимизации (1.30) в условиях работы [488]:

$$f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle \to \min_{x \in \mathbb{R}^n},$$

где $A = \|A_{ij}\|_{i,j=1}^n \succ 0$ и $1 \leqslant A_{ij} \leqslant 2$ при $i, j = 1, \ldots, n$. Из последнего условия следует, что

$$L = \lambda_{\max}(A) \geqslant \lambda_{\max}(1_n 1_n^T) = n,$$

поэтому оценка общего времени работы МЛК (оптимального, с точностью до числового множителя, метода для данного класса задач) будет составлять

$$\underbrace{\mathrm{O}(n^2)}_{\substack{\text{стоимость}\\ \text{итерации}}} \cdot \underbrace{\mathrm{O}\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)}_{\substack{\text{число}\\ \text{итераций}}} = \mathrm{O}\left(\frac{n^{5/2}R}{\varepsilon^{1/2}}\right).$$

При этом если использовать покомпонентную рандомизацию с вероятностью выбрать i-орт $p_i \sim \sqrt{L_i}$, где $L_i = A_{ii}$ — константа Липшица i-й компоненты градиента вдоль i-орта, то в оценке числа итераций вместо $\sqrt{L} \geqslant \sqrt{n}$ можно ставить

$$\sqrt{\bar{L}} = \frac{1}{n}\sum_{i=1}^n \sqrt{L_i} \leqslant \sqrt{2},$$

т. е. число итераций согласно замечанию 1 (см. также [21, 488]) будет равно

$$\mathrm{O}\left(n\sqrt{\frac{\bar{L}R^2}{\varepsilon}}\right) = \mathrm{O}\left(\frac{nR}{\varepsilon^{1/2}}\right).$$

При этом стоимость итерации теперь будет составлять $\mathrm{O}(n)$. Действительно, если $\tilde{x} = x + he_i$, где e_i — i-орт и уже посчитано Ax, то

$$A\alpha\tilde{x} = \alpha Ax + \alpha h A^{\langle i\rangle},$$

где α — заданное число, может быть посчитано за время $\mathrm{O}(n)$. Отсюда с учётом указания к упражнению 1.3 следует, что и итерацию можно осуществить за время $\mathrm{O}(n)$. Таким образом, общее время работы ПМЛК с $\beta = 1/2$ можно оценить следующим образом:

$$\underbrace{\mathrm{O}(n)}_{\substack{\text{стоимость}\\ \text{итерации}}} \cdot \underbrace{\mathrm{O}\left(n\sqrt{\frac{\bar{L}R^2}{\varepsilon}}\right)}_{\substack{\text{число}\\ \text{итераций}}} = \mathrm{O}\left(\frac{n^2R}{\varepsilon^{1/2}}\right) \ll \mathrm{O}\left(\frac{n^{5/2}R}{\varepsilon^{1/2}}\right).$$

Отметим, что если «собирать» из компонент градиента полный градиент и использовать обычный быстрый градиентный метод, то в последней оценке вместо $\bar{L}$ следовало бы писать просто L. Таким образом, в общем случае в задачах с неполной информацией о градиенте не следует использовать полноградиентные методы «напрямую» — путём полного восстановления градиента по частично наблюдаемой информации.

Этот же вывод (уже по другой причине) можно сделать и для спусков по направлению и безградиентных методов [9, 16, 267, 268].

Кажется, что получается какое-то противоречие с оптимальностью МЛК для класса гладких выпуклых задач и противоречие с рядом тезисов, которые приводились в пособии ранее. На самом деле никаких противоречий нет [74].

1. Во-первых, нижние оценки были получены на классе детерминированных методов. Впрочем, введение рандомизации принципиально не меняет нижние оценки.
2. Во-вторых, МЛК оптимален на классе всевозможных гладких выпуклых задач с ограниченной константой Липшица градиента. Выше же был рассмотрен более узкий класс задач — квадратичной оптимизации. Но даже при таком сужении нижние оценки принципиально не изменятся (см. упражнение 1.3). Более важным и ограничительным было предположение на элементы матрицы A, которое привело к тому, что след матрицы A имеет тот же порядок, что и наибольшее собственное значение. Именно эта «асимметричность» постановки задачи и определила успешность использования рандомизации.
3. В-третьих, МЛК оптимален на классе задач по числу вычислений градиента функционала (в нашем случае — по числу произведений матрицы на вектор), а не по времени работы (общей трудоёмкости). По времени работы не существует теории нижних оценок, и вряд ли в ближайшее время можно ожидать её появления.
4. Наконец, ранее отмечалось, что если есть возможность считать градиент, например, с помощью автоматического дифференцирования, то следует именно градиент и использовать в методе, а не восстанавливать компоненты градиента по посчитанным значениям функции. Однако в последнем примере используется не процедура расчёта значений функции, а процедура их пересчёта. Действительно, вычислив самый первый раз $f(x)$ за время $O(n^2)$, дальше можно уже пересчитывать $f(A\alpha\tilde{x})$, где $\tilde{x} = x + he_i$, передавая Ax с прошлого запуска, за время $O(n)$. Для осуществления шага нужно посчитать только одну (случайно выбранную) компоненту градиента, что может быть приближённо сделано за два пересчёта значения функции, т. е. за время $O(n)$.

♦ Заметим, что аналогичный результат (с точностью до логарифмического множителя) можно было бы получить на основе неускорен-

ного покомпонентного метода с $\gamma = 1$ [186, п. 6.4.2], [469] и техники каталист, см. замечание 3.3. При этом для такого ускоренного покомпонентного метода (в отличие от всех других, известных нам [21]) уже можно обеспечить учёт разреженности матрицы A в оценке стоимости итерации при минимизации функции [352]

$$\mu \ln\Big(\sum_{k=1}^{m} \exp\Big(\frac{\langle A_k, x\rangle}{\mu} \Big) \Big).$$

Функции такого типа возникают, например, при решении двойственной задачи к энтропийно-регуляризованной транспортной задаче ЛП, см. упражнение 3.9 и замечание 4.1.

Отмеченную (двойственную) функцию также можно минимизировать с помощью тензорных методов [193, 266, 475] (см. замечание 3.3 и текст после него). При этом для восстановления решения прямой задачи ввиду замечания 5.3 и упражнения 4.2 в качестве критерия следует выбирать норму градиента. Регуляризация двойственной функции (см. замечание 4.1 и упражнение 4.4) позволяет добиться оптимальной скорости сходимости [266] (см. также работу [315], содержащую нижние оценки) по критерию нормы градиента соответствующих тензорных методов без потери логарифмического множителя в скорости сходимости, как это происходило для методов первого порядка, см. замечание 5.3. ♦

Может показаться, что выше слишком много внимания было уделено, на первый взгляд, довольно незначительной оговорке о возможности в специальных (асимметричных) случаях ускорять время работы методов в $\sim \sqrt{n}$ раз за счёт покомпонентной рандомизации используемых методов.

Одна из причин такого внимания к данному примеру связана с так называемым «препроцессингом». А именно, в описанном выше подходе при рандомизации $p_i \sim \sqrt{L_i}$ явно использовалась дополнительная информация о структуре задачи. Эту информацию несложно было получить до начала работы метода. Хотя в данном случае основной эффект достигался в первую очередь за счёт самого факта рандомизации, а не за счёт её специфики [21], в общем случае следует признать удачными в оптимизационной практике приёмы типа диагонального шкалирования (предобуславливание), пришедшие из вычислительной линейной алгебры [37, 95, 304, 499, 585], и их рандомизированные варианты типа описанного выше. Особенно популярны сейчас (в связи с приложениями к анализу данных) адаптивные варианты

таких методов, в которых по ходу работы метода предпринимается попытка без больших усилий (т. е. не как в методе Ньютона) улучшить обусловленность задачи. К таким методам можно отнести уже упоминавшиеся ранее методы с процедурой растяжения пространства [90, 104] и предобуславливание, описанное в упражнении 3.11 и указании к упражнению 5.9. Также популярными сейчас являются *квазиньютоновские методы* [495] (см. также замечание 3) и методы типа *AdaGrad* [40, 230, 258, 533].

Однако главная причина такого внимания к покомпонентным методам описана ниже. Исследования последних лет показывают (см., например, [21] и цитированную там литературу), что именно покомпонентная рандомизация (для прямой или двойственной задачи) лежит в основе большей части современных подходов к решению задач оптимизации, приходящих из анализа данных [547]. В частности, различные варианты метода рандомизации суммы (см. замечание 1) можно понимать как варианты покомпонентных рандомизаций для двойственной задачи.

Пример (типичная задача анализа данных [21, 120, 179, 547, 550]). Рассмотрим задачу выпуклой оптимизации:

$$\sum_{k=1}^{m} f_k(A_k^T x) + g(x) \to \min_{x \in Q},$$

где $g(x) = \sum_{i=1}^{n} g_i(x_i)$ (это условие нужно только для того, чтобы можно было эффективно построить двойственную задачу, — см. ниже), а Q — множество простой структуры. Предполагаем, что трудоёмкость вычисления $f_k'(z_k)$ равна $O(1)$ и что для всех $k = 1, \ldots, m$ и всех допустимых w, v выполнено неравенство

$$|f_k'(w) - f_k'(v)| \leqslant L|w - v|.$$

Функция $g(x)$ предполагается сильно выпуклой в p-норме с константой μ_p. Вводя матрицу $A = [A_1, \ldots, A_m]^T$ и вспомогательный вектор $z = Ax$, мы можем переписать эту задачу в «раздутом» пространстве $\tilde{y} = (x, z)$ как задачу типа проектирования на аффинное многообразие [4, 135, 255], см. также формулу (4.7). Эта задача проектирования решается путём перехода к двойственной задаче. Опишем далее довольно общий способ построения двойственной задачи:

$$\min_{x \in Q} \left\{ \sum_{k=1}^{m} f_k(A_k^T x) + g(x) \right\} = \min_{\substack{x \in Q \\ z = Ax}} \left\{ \sum_{k=1}^{m} f_k(z_k) + g(x) \right\} =$$

$$= \min_{\substack{x\in Q\\ z=Ax,\, z'}} \max_{y} \left\{ \langle z - z', y\rangle + \sum_{k=1}^{m} f_k(z'_k) + g(x) \right\} =$$

$$= \max_{y\in\mathbb{R}^m} \left\{ -\max_{\substack{x\in Q\\ z=Ax}} \{\langle -z, y\rangle - g(x)\} - \max_{z'} \left\{ \langle z', y\rangle - \sum_{k=1}^{m} f_k(z'_k) \right\} \right\} =$$

$$= \max_{y\in\mathbb{R}^m} \left\{ -\max_{x\in Q} \left\{ \langle -A^T y, x\rangle - g(x) \right\} - \sum_{k=1}^{m} \max_{z'_k} \left\{ z'_k y_k - f_k(z'_k) \right\} \right\} =$$

$$= \max_{y\in\mathbb{R}^m} \left\{ -g^*(-A^T y) - \sum_{k=1}^{m} f_k^*(y_k) \right\} =$$

$$= -\min_{y\in\mathbb{R}^m} \left\{ g^*(-A^T y) + \sum_{k=1}^{m} f_k^*(y_k) \right\}.$$

В описанную схему погружаются, скажем, следующие задачи [135]:

1) $\dfrac{L}{2}\|Ax-b\|_2^2 + \underbrace{\dfrac{\mu}{2}\|x-x_g\|_2^2}_{g(x)} \to \min\limits_{x\in\mathbb{R}^n}$ (Ridge regression);

2) $\dfrac{L}{2}\|Ax-b\|_2^2 + \underbrace{\mu\sum\limits_{k=1}^{n} x_k \ln x_k}_{g(x)} \to \min\limits_{x\in S_n(1)}$ (Minimal mutual information model).

Константа Липшица производной функции $f_k(z) = \dfrac{L}{2}(z_k - b_k)^2$ равна L, константы сильной выпуклости функции $g(x)$ (считаются в разных нормах: 1) в 2-норме, 2) в 1-норме) также одинаковы в обоих случаях и равны μ. Для приведённых выше задач получим следующие двойственные задачи:

1) $\dfrac{1}{2\mu}\left(\|x_g - A^T y\|_2^2 - \|x_g\|_2^2\right) + \dfrac{1}{2L}\left(\|y+b\|_2^2 - \|b\|_2^2\right) \to \min\limits_{y\in\mathbb{R}^m}$;

2) $\mu \ln\left(\sum\limits_{i=1}^{n} \exp\left(\dfrac{[-A^T y]_i}{\mu}\right)\right) + \dfrac{1}{2L}\left(\|y+b\|_2^2 - \|b\|_2^2\right) \to \min\limits_{y\in\mathbb{R}^m}$.

В общем случае можно утверждать, что $\sum_{k=1}^{m} f_k^*(y_k)$ (композитный член в двойственной задаче, см. пример 3.1) является сильно выпуклым в 2-норме с константой сильной выпуклости, равной L^{-1}. Для $g^*(-A^T y)$ можно оценить константу Липшица градиента в 2-норме (см. конец § 2, начало § 4, а также указание к упражнению 4.8):

$$\frac{1}{\mu}\max_{\|y\|_2\leqslant 1,\, \|x\|_p\leqslant 1} \langle A^T y, x\rangle^2 = \frac{1}{\mu}\max_{\|x\|_p\leqslant 1} \|Ax\|_2^2 = \frac{1}{\mu}\begin{cases} 1)\ \lambda_{\max}(A^T A), \\ 2)\ \max\limits_{j=1,\ldots,n} \|A^j\|_2^2 \end{cases}$$

и получить следующую оценку сверху на константы Липшица всех частных производных $g^*(-A^Ty)$ [21]:

$$\frac{1}{\mu}\max_{\|y\|_1\leqslant 1,\,\|x\|_p\leqslant 1}\langle A^Ty,x\rangle^2=\frac{1}{\mu}\max_{\|x\|_p\leqslant 1}\|Ax\|_\infty^2=\frac{1}{\mu}\begin{cases}1)\ \max\limits_{i=1,\ldots,m}\|A_i\|_2^2,\\ 2)\ \max\limits_{\substack{i=1,\ldots,m\\ j=1,\ldots,n}}|A_{ij}|^2.\end{cases}$$

Для ПМЛК с $\beta=0$ из замечания 2, применённого к двойственной задаче (в случае, когда матрица A плотно заполненная), имеем следующие оценки общего времени работы (трудоёмкости):

1) $T_1=\widetilde{O}\left(n\cdot m\sqrt{\frac{L\max_{i=1,\ldots,m}\|A_i\|_2^2}{\mu}}\right)$;

2) $T_2=\widetilde{O}\left(n\cdot m\sqrt{\frac{L\max_{i,j}|A_{ij}|^2}{\mu}}\right)$.

Если теперь посмотреть на исходную прямую задачу (с $L:=L/m$):

$$\frac{1}{m}\sum_{k=1}^m f_k(A_k^Tx)+g(x)\to\min_{x\in Q}$$

и оценить трудоёмкость оптимальных методов согласно оценкам, приведённым в конце замечания 1, то получим уже анонсированное соответствие оптимальных оценок с полученными только что оценками трудоёмкости ПМЛК с $\beta=0$ (при $L:=L/m$). Действительно, с учётом того, что константы Липшица градиентов функций $f_k(A_k^Tx)$, посчитанные в нормах, соответствующих норме, в которой сильно выпукл композит прямой задачи, равномерно (по $k=1,\ldots,m$) оцениваются следующим образом:

$$1)\ L\max_{i=1,\ldots,m}\|A_i\|_2^2;\quad 2)\ L\max_{i,j}|A_{ij}|^2,$$

а сложность вычисления $\nabla f_k(A_k^Tx)$ равна $O(n)$, в типичном случае получаем

1) $\widetilde{T}_1=\widetilde{O}\left(n\cdot\left(m+\sqrt{m\frac{L\max_{i=1,\ldots,m}\|A_i\|_2^2}{\mu}}\right)\right)=\widetilde{O}\left(n\cdot m\sqrt{\frac{\frac{L}{m}\max_{i=1,\ldots,m}\|A_i\|_2^2}{\mu}}\right)$;

2) $\widetilde{T}_2=\widetilde{O}\left(n\cdot\left(m+\sqrt{m\frac{L\max_{i,j}|A_{ij}|^2}{\mu}}\right)\right)=\widetilde{O}\left(n\cdot m\sqrt{\frac{\frac{L}{m}\max_{i,j}|A_{ij}|^2}{\mu}}\right)$.

Таким образом, имеет место полное соответствие (с точностью до опущенных в рассуждениях логарифмических множителей). Инте-

ресно заметить, что для первой задачи здесь можно сполна использовать разреженность матрицы A [21]. Более того, эту задачу (также с полным учётом разреженности) можно решать и прямым ПМЛК с $\beta = 0$. Соответствующие оценки согласно замечанию 2 имеют вид (снова возвращаемся к исходному пониманию параметра L)

$$T_1^{\text{прям}} = \widetilde{\mathrm{O}}\left(s \cdot n \sqrt{\frac{L \max_{j=1,\ldots,n} \|A^j\|_2^2}{\mu}}\right),$$

$$T_1^{\text{двойств}} = \widetilde{\mathrm{O}}\left(\tilde{s} \cdot m \sqrt{\frac{L \max_{i=1,\ldots,m} \|A_i\|_2^2}{\mu}}\right) = \widetilde{\mathrm{O}}\left(sn \sqrt{\frac{L \max_{i=1,\ldots,m} \|A_i\|_2^2}{\mu}}\right),$$

где s — среднее число ненулевых элементов в столбцах матрицы A, а $\tilde{s}$ — в строках.

В действительности обе эти оценки оказываются завышенными[7]. Более аккуратные рассуждения позволяют заменить максимум в этих оценках на некоторые средние. Скажем, в транспортных приложениях [32, гл. 1], [135] матрица A не просто разреженная, но ещё и битовая (состоит из нулей и единиц). В таком случае приведённые оценки переписываются следующим образом:

$$T_1^{\text{прям}} = \widetilde{\mathrm{O}}\left(sn \sqrt{\frac{Ls}{\mu}}\right), \quad T_1^{\text{двойств}} = \widetilde{\mathrm{O}}\left(sn \sqrt{\frac{L\tilde{s}}{\mu}}\right).$$

Отсюда можно сделать довольно неожиданный вывод [135]: при $m \ll n$ следует использовать прямой ПМЛК, а в случае $m \gg n$ — двойственный. Первый случай соответствует приложениям к изучению больших сетей (компьютерных, транспортных). Второй случай соответствует задачам, приходящим из анализа данных.

В заключение отметим, что описанный в данном примере подход используется при построении эффективных асинхронных распределённых алгоритмов [339, 341], т. е. алгоритмов, не требующих синхронизации распределённых устройств. ■

Несмотря на большую активность в последние годы в направлении разработки рандомизированных методов выпуклой оптимизации, здесь по-прежнему остаются открытые вопросы, например вопрос о возможности построения адаптивных ускоренных методов. Неускоренный блочно-покомпонентный градиентный спуск с адаптивным

[7] Это легко усмотреть из способа рассуждений, в котором мы заменяем константы Липшица частных производных на худшую из них. Отметим, что описанное огрубление также упрощает использование оценки скорости сходимости ПМЛК с $\beta = 0$ из замечания 2.

подбором констант Липшица соответствующих блоков компонент градиента был предложен в работе [469]. Полностью адаптивный ускоренный аналог до сих пор не известен. Известны лишь частично адаптивные конструкции, см., например, [281, 352].

♦ Отметим, что описанное выше направление в его современном варианте появилось во многом под влиянием двух препринтов Ю. Е. Нестерова, подготовленных в 2010–2011 гг. Результаты, приведённые в этих препринтах, впоследствии вошли в две статьи [469, 487]. П. Рихтарик (постдок Ю. Е. Нестерова) активно взялся за развитие отмеченных идей (статей) Ю. Е. Нестерова, особенно в части покомпонентных методов, находя им (их параллельным и распределённым вариантам) всевозможные приложения, особенно в задачах огромных размеров (*Big Data*), приходящих из анализа данных [524]. ♦

Для рандомизированных методов в случае возможности вычислять значение оптимизируемой функции $f(x)$ из результатов о сходимости в среднем

$$E[f(x^N)] - f(x_*) \leqslant \varepsilon$$

по неравенству Маркова следует, что

$$P\left(f(x^N) - f(x_*) \geqslant 2\varepsilon\right) \leqslant \frac{E(f(x^N) - f(x_*))}{2\varepsilon} \leqslant \frac{1}{2}.$$

Запуская параллельно $m = \lceil \log_2(\sigma^{-1}) \rceil$ независимых траекторий метода и выбирая $\widehat{x}^N = \arg\min_{k=1,\ldots,m} f(x^{N,k})$, получим [21]

$$P\left(f(\widehat{x}^N) - f(x_*) \geqslant 2\varepsilon\right) \leqslant \sigma.$$

Некоторые тонкости возникают для задач условной оптимизации [32, п. 4.4], [158], в которых точка x^N может быть недопустимой (не удовлетворять ограничениям) и, как следствие, использовать неравенство Маркова нельзя ввиду отсутствия гарантий, что будет выполнено условие $f(x^N) - f(x_*) \geqslant 0$. Аналогичные результаты об увеличении числа вычислений (в данном случае числа итераций) в $\sim \ln(\sigma^{-1})$ раз можно получить, не предполагая возможности вычислять значения $f(x)$ (что затруднительно, например, для обычных задач стохастической оптимизации [553]), но предполагая, что стохастические (суб-)градиенты несмещённые и имеют финитный носитель. Если хвосты стохастических (суб-)градиентов субгауссовские, то вместо $\sim \ln(\sigma^{-1})$ следует писать $\sim \ln^2(\sigma^{-1})$. В общем случае см., например, [20, 44, 233, 271, 366, 410, 553]. Отметим, что в ряде случаев возможность параллелизации вычислений без доступа к значениям $f(x)$ можно сохранить

[44, 302], используя просто немного другой способ формирования $\widehat{x}^N$ (похожие идеи используются и в подходе local SGD [383, 390, 565, 606]):

$$\widehat{x}^N = \frac{1}{m}\sum_{k=1}^{m} x^{N,k}.$$

Если рандомизированный метод линейно сходится в среднем, см., например, [124, 413], т. е. $E[f(x^{N(\varepsilon)})] - f(x_*) \leqslant \varepsilon$ и $N(\varepsilon) \sim \ln \varepsilon^{-1}$, то для получения точных оценок вероятностей больших уклонений также можно использовать грубое неравенство Маркова [135]:

$$P\big(f(x^{N(\varepsilon\sigma)}) - f(x_*) \geqslant \varepsilon\big) \leqslant \sigma \frac{E[f(x^{N(\varepsilon\sigma)})] - f(x_*)}{\varepsilon\sigma} \leqslant \sigma.$$

♦ Пусть $(f_\delta(x); \langle \nabla f_\delta(x), y - x\rangle)$ — (δ, L)-модель функции $f(x)$ в точке x (относительно нормы $\|\ \|$), см. § 3. Пусть (*субгауссовские хвосты* у модели стохастического градиента)[8]

$$E_\xi[\nabla_x f_\delta(x,\xi)] \equiv \nabla f_\delta(x), \quad E_\xi\left[\exp\left(\frac{\|\nabla_x f_\delta(x,\xi) - \nabla f_\delta(x)\|_*^2}{D}\right)\right] \leqslant \exp(1).$$

Тогда имеют место следующие неравенства [96], [233, гл. 7], [320, 362], [368, теорема 2.1], [409, лемма 2]:

$$f_\delta(y) \leqslant f(y) \leqslant f_\delta(y) + \delta;$$

$$P\Big(f_\delta(y) \leqslant f_\delta(x) + \Big\langle \overset{r}{\nabla}_x f_\delta(x, \{\xi^l\}_{l=1}^r), y - x\Big\rangle + \frac{2L}{2}\|y - x\|^2 + \\ + \frac{3(2\chi + 4\Omega\chi + 2\Omega^2)}{2Lr} + \delta\Big) \geqslant 1 - \exp\Big(-\frac{\Omega^2}{3}\Big),$$

[8] Условия субгауссовости хвостов всегда можно добиться с помощью «bias variance trade off»: введения смещения (bias) в стохастический градиент за счёт обнуления стохастического градиента вне некоторого шара. Радиус шара выбирается таким образом, чтобы смещение в итоговой оценке давало вклад, не больший, чем желаемая точность решения задачи. Субгауссовская дисперсия (variance), входящая в условие субгауссовости для так «обрезанного» стохастического градиента, будет заметно больше его обычной дисперсии. Хочется учесть это обстоятельство. Поэтому вместо неравенства Азума — Хёфдинга (приведённого далее) следует использовать более тонкое неравенство Бернштейна — Фридмана. Если так действовать, то можно даже в случае тяжёлых хвостов (в предположении конечной дисперсии исходного стохастического градиента) получить экспоненциальную концентрацию в оценках скорости сходимости соответствующих методов для гладких задач выпуклой оптимизации. Для неускоренных методов это сделано в работе [70], а для ускоренных с помощью другой техники аналогичный результат был получен в работе [225]. По-видимому, техника работы [70] также позволяет распространить результаты этой работы на ускоренные методы (недавно это было сделано [306, 613]) и на гладкие монотонные вариационные неравенства, см. замечание 5.1.

где $\chi = 1$ для $\|\ \| = \|\ \|_2$ и $\chi = \min\{q-1, 2\ln n\}$ для $\|\ \| = \|\ \|_p$, $\frac{1}{p} + \frac{1}{q} = 1$, $p \in [0, \infty]$;

$$P\left(\sum_{k=0}^{N-1} c_{k+1} \left\langle \nabla_x f_\delta\left(x^k, \{\xi^{k+1,l}\}_{l=1}^{r^{k+1}}\right) - \nabla f_\delta(x^k), x - x^k \right\rangle \right.$$
$$\left. \leqslant \sqrt{3DR^2\Omega \sum_{k=0}^{N-1} \frac{c_{k+1}^2}{r_{k+1}}} \right) \geqslant 1 - \exp(-\Omega),$$

где по предположению величины c_{k+1}, r^{k+1} и x^k могут зависеть только от $\left\{\{\xi^{t+1,l}\}_{l=1}^{r^{t+1}}\right\}_{t=0}^{k-1}$, а $\|x - x^k\|_2 \leqslant R$ при $k = 0, \ldots, N-1$, $\Omega \geqslant 0$. Последнее неравенство для вероятности больших уклонений обычно называют *неравенством концентрации* (*меры*) *Азума — Хёфдинга* [178, 233, 553].

Приведённые выше неравенства в сочетании с упражнением 3.7 позволяют получить оптимальные методы решения гладких выпуклых задач стохастической оптимизации в модельной общности [43]. ♦

В описанной выше общности (см., в частности, работы [26, 43, 272, 569]) полученная линейка методов уже будет практически полностью покрывать основной арсенал методов первого порядка (и ниже), использующихся в современных приложениях для решения задач выпуклой оптимизации большого размера.

- В частности, описанные выше в пособии подходы (в особенности прямодвойственные универсальные ускоренные методы решения седловых задач, задач выпуклой оптимизации с оракулом, выдающим модель функции, и их (блочно) покомпонентные варианты) позволяют строить методы, работающие по наилучшим известным сейчас теоретическим оценкам общего времени работы (трудоёмкости) практически для всех известных нам классов задач (структурной) выпуклой оптимизации больших размеров.

Большое число исследований во всём мире сейчас сосредоточено на изучении (переборе) конкретных способов сочетания описанных выше приёмов с целью определения их наилучшего сочетания для изучаемого класса задач, как правило, возникающего в одном из актуальных приложений. Отметим в этой связи упражнения 3.7, 5.8 и замечание 3.3, «ускоряющие» почти все оценки, приведённые в пособии, и упражнение 5.5, распространяющее действие разобранных в пособии конструкций с задач выпуклой оптимизации на множествах простой структуры на общие задачи выпуклой оптимизации (с аффинными ограничениями вида равенств и выпуклыми ограничениями

вида неравенств). Отметим также, что многие популярные на практике приёмы типа *альтернативных направлений*, *метода штрафных функций*, *метода модифицированной функции Лагранжа*, *ADMM* и др. [13, 52, 56, 86, 92, 157, 160, 161, 226, 403, 407, 408, 437, 495, 501, 611] и их ускоренные варианты — ускоренный метод альтернативных направлений, AADMM [284, 303, 323, 371, 384, 437, 501] — не гарантируют в общем случае лучших теоретических оценок, чем те, которые могут быть получены при описанных в пособии подходах. Улучшения могут быть только в негладком случае за счёт проксимальной природы ряда описанных процедур, например метода модифицированной функции Лагранжа [584], см. также § 3, упражнение 4.3 и замечание 4.3. Но сполна этим можно воспользоваться, как правило, только в случае организации распределённых вычислений, см., например, метод ADMM [179, 501, 611].

Однако не следует думать, что этим уже исчерпываются современные численные методы выпуклой оптимизации и способы их исследования. Выше мало обсуждались вопросы о стоимости итерации, завязанные на автоматическое дифференцирование [49, 159, 495] и возможность быстрого пересчёта значений функции и компонент градиента [32, 491]. В этой связи достаточно привести несколько примеров [32, п. 1.5.2, 4, 5.1], [129, 130], [186, п. 3.3], [219, 346, 366, 423, 484, 491], см. также упражнение 1.6. Все эти примеры ярко демонстрируют, что на практике за счёт дешёвых итераций быстрее могут работать методы, которые далеко не оптимальны с точки зрения числа итераций. Также совсем не рассматривались реальные приложения, например к задачам оптимизации в гильбертовых пространствах, в которых оракул типично зашумлён [14, 48, 49, 292, 507]. Наконец, почти ничего не было сказано о многих других методах и подходах, которые часто используются при решении практических задач умеренных размеров [282], в частности о квазиньютоновских методах [495] и методах второго порядка (ньютоновские методы), интерес к которым в последние годы резко возрос [25, 36, 49, 75, 76, 112, 119, 144, 150, 197, 217, 297, 317, 318, 374, 375, 436, 450, 468, 473–477, 480, 597, 608].

Замечание 3 (квазиньютоновские методы [232], [495, гл. 6]). В замечании 1.2 приводится геометрическая интерпретация градиентного спуска, в основу которой положена замена исходной функции параболоидом вращения, касающимся её графика в текущей точке. Точка, доставляющая минимум параболоиду, принимается за новое положение метода. В замечании 1.4 рассматривается наискорейший

спуск, заключающийся в подборе кривизны параболоида с помощью решения вспомогательной задачи одномерной минимизации. В *методе Ньютона* вместо параболоида вращения строится квадратичная аппроксимация оптимизируемой функции (на основе доступного гессиана), однако это приводит к необходимости решения на каждом шаге более сложной задачи — минимизации квадратичной формы (1.30). Естественно возникает идея построения какого-то «промежуточного» метода, с одной стороны, не требующего вычисления (и тем более обращения) гессиана, а с другой стороны, всё-таки пытающегося как-то аппроксимировать гессиан исходя из накопленной информации первого порядка (градиентов). В основу *квазиньютоновских методов* положен следующий общий принцип построения квадратичной аппроксимации: квадратичная аппроксимация должна касаться графика оптимизируемой функции в текущей точке и иметь с ним одинаковые градиенты в точке с предыдущего шага (secant equation). Существует много различных способов удовлетворить этим условиям. У этих способов есть различные интерпретации, среди которых отметим понимание квазиньютоновских методов как *методов переменной метрики* [86, п. 2, § 3, гл. 3] (варианта метода сопряжённых градиентов). Наиболее интересными в практическом плане являются способы, которые требуют не более чем квадратичной (по размерности пространства) трудоёмкости и памяти. Среди таких способов наиболее удачно себя зарекомендовал способ, приводящий в итоге к методу *BFGS*:

$$h_k = \arg\min_{h\in\mathbb{R}} f\left(x^k - hH_k\nabla f(x^k)\right), \quad x^{k+1} = x^k - h_k H_k \nabla f(x^k),$$

$$H_{k+1} = H_k + \frac{H_k\gamma_k\delta_k^T + \delta_k\gamma_k^T H_k}{\langle H_k\gamma_k, \gamma_k\rangle} - \beta_k \frac{H_k\gamma_k\gamma_k^T H_k}{\langle H_k\gamma_k, \gamma_k\rangle},$$

где

$$\beta_k = 1 + \frac{\langle \gamma_k, \delta_k\rangle}{\langle H_k\gamma_k, \gamma_k\rangle},$$

$$\gamma_k = \nabla f(x^{k+1}) - \nabla f(x^k), \quad \delta_k = x^{k+1} - x^k, \quad H_0 = I.$$

В отличие от сопряжённых градиентов (см. замечание 1.6), в BFGS не обязательно точно осуществлять вспомогательную одномерную оптимизацию. В целом BFGS оказался наиболее устойчивым (к вычислительным погрешностям) вариантом квазиньютоновских методов. Геометрия квазиньютоновских методов (см. замечание 1.2) близка (но не идентична!) геометрии субградиентных методов с процедурой растяжения пространства [86, п. 4, § 4, гл. 5] (метод эллипсоидов, методы Шора). Рисунок 4 (с. 43)демонстрирует типичное «пилообразное»

поведение градиентных спусков в окрестности минимума для плохо обусловленных задач. Из этого рисунка видно, что направления γ_k и δ_k «подсказывают» направления растяжения/сжатия и позволяют адаптивно улучшать обусловленность задачи, правильно аккумулируя собранную информацию в H_k.

В теоретическом плане про квазиньютоновские методы на данный момент (как и 30–40 лет назад) известно не так уж и много (см., например, [76, п. 1.3.1], [232], [495, гл. 6]): глобальная скорость сходимости для гладких задач выпуклой оптимизации в общем случае не выше, чем у обычных (неускоренных) градиентных методов (во всяком случае, только это пока удалось установить), а локальная скорость сходимости в случае невырожденного минимума *сверхлинейная*, т. е. быстрее, чем линейная. В работах [528, 529] был предложен такой вариант квазиньютоновского метода, для которого в случае невырожденного минимума была доказана оценка локальной скорости сходимости вида $\|x^k - x_*\|_2 \sim c^{k^2}$, где $c \in (0, 1)$ — некоторая константа, связанная с числом обусловленности задачи.

Основным ограничением по использованию квазиньютоновских методов является необходимость хранения и обновления плотной квадратной матрицы H_k, что требует (в отличие от того, что имеет место для методов типа сопряжённых градиентов) квадратичной памяти и квадратичного времени независимо от разреженности задачи. Это обстоятельство существенно ограничивает возможности использования таких методов для задач оптимизации с десятками тысяч переменных и более. Однако на практике используют в основном варианты таких методов *с ограниченной памятью*, см., например, метод *LBFGS* [495, п. 7.2]. В этом случае в памяти хранится не матрица H_k, а векторы, её порождающие. Проблема, однако, тут в том, что с ростом k размер этой памяти линейно растёт. Поэтому обычно последовательности векторов $\{\gamma_l\}_{l=0}^k$ и $\{\delta_l\}_{l=0}^k$ хранят только с q последних итераций (q — глубина памяти) и при этом полагают $H_{k-q} = I$. На практике q часто выбирают совсем небольшим: $q \simeq 3-5$. Впрочем, нам на практике встречались примеры невыпуклых задач обучения, в которых наилучшим образом сходился вариант LBFGS с $q = 900$ [299].

Есть основания полагать, что в ближайшее время именно в данной области могут появиться наиболее интересные результаты, объясняющие высокую эффективность на практике таких методов, как, например, LBFGS [495, гл. 7], и всевозможных рандомизированных вариантов квазиньютоновских методов [177, 312, 533]. ■

В заключение всё же приведём сопоставительный анализ методов первого порядка (градиентных методов) и методов более высокого порядка, которые могут использоваться для решения задач выпуклой оптимизации умеренных размеров ($n \leqslant 10^4$), в условиях отсутствия шума на классе достаточно гладких, т. е. удовлетворяющих условиям

$$\|\nabla^r f(y) - \nabla^r f(x)\|_2 \leqslant M_r \|y - x\|_2,$$
$$x, y \in \mathbb{R}^n, \quad M_r \leqslant \infty, \quad r = 0, 1, 2, \ldots,$$

μ-сильно выпуклых в 2-норме задач, где $\mu \geqslant 0$. Условие $x, y \in \mathbb{R}^n$ можно заменить условием $x, y \in \{z \in \mathbb{R}^n : f(z) \leqslant f(x^0)\}$. При $r = 0, 1$ (и при $r \geqslant 2$ для оптимального тензорного метода из замечания 3.3, см. также [25, 189, 476]) последнее условие, в свою очередь, можно улучшить, см., например, (2.19). Условие липшицевости $\nabla^r f(x)$ можно заменить условием гёльдеровости $\nabla^r f(x)$ с параметром $\nu \in [0, 1]$ [316]. При этом приведённые далее оптимальные оценки можно распространить (если параметр ν известен) и на этот случай [559].

♦ Заметим, что $\nabla^r f(y)$ — тензор ранга r. Поэтому следует пояснить, что понимается под 2-нормой в левой части данного неравенства. Ограничимся случаем $r = 2$, тогда

$$\nabla^2 f(x) = \left\| \frac{\partial^2 f(x)}{\partial x_i \partial x_j} \right\|_{i,j=1}^n,$$

$$\|\nabla^2 f(y) - \nabla^2 f(x)\|_2 = \sup_{\|h_1\|_2 \leqslant 1} \sup_{\|h_2\|_2 \leqslant 1} \left\langle \left(\nabla^2 f(y) - \nabla^2 f(x)\right)[h_1], h_2 \right\rangle =$$
$$= \sup_{\|h_1\|_2 \leqslant 1} \sup_{\|h_2\|_2 \leqslant 1} \left\langle \left(\nabla^2 f(y) - \nabla^2 f(x)\right) h_1, h_2 \right\rangle = \sup_{\|h\|_2 \leqslant 1} \left\langle \left(\nabla^2 f(y) - \nabla^2 f(x)\right) h, h \right\rangle =$$
$$= \max \left\{ \lambda_{\max}\left(\nabla^2 f(y) - \nabla^2 f(x)\right), \left|\lambda_{\min}\left(\nabla^2 f(y) - \nabla^2 f(x)\right)\right| \right\}.$$

В общем случае см. [150, 473]. Отметим также, что при $r = 0$ мы имеем $\nabla^0 f(x) = f(x)$, а $\| \ \|_2 = | \ |$. В связи с последним замечанием стоит упомянуть, что в действительности под M_0 можно понимать меньшую константу (а именно L_0 — см., например, формулу (2.4)), которая только в худшем случае совпадает с введённой здесь [479]. Аналогичное замечание имеет место и для методов 2-го порядка [480] (и, вероятно, более высокого порядка). «Правильный» метод p-го порядка ($p \geqslant 1$) на первых итерациях «осуществляет» желаемую редукцию (уменьшение) констант гладкости $\{M_r\}_{r=0}^{p-1}$ за счёт попадания в нужную область сходимости метода. При этом часто достаточно одной (первой) итерации [477, 479, 480]. ♦

Для класса методов, в которых на каждой итерации разрешается не более чем O(1) раз обращаться к оракулу (подпрограмме) за значениями $\nabla^r f(x)$, $r \leqslant 1$, оценка числа итераций, необходимых для достижения точности ε (по функции), будет иметь вид

$$\mathrm{O}\left(\min\left\{n\ln\left(\frac{\Delta f}{\varepsilon}\right);\ \frac{M_0^2R^2}{\varepsilon^2},\ \left(\frac{M_1R^2}{\varepsilon}\right)^{1/2};\ \frac{M_0^2}{\mu\varepsilon},\ \left(\frac{M_1}{\mu}\right)^{1/2}\cdot\left\lceil\ln\left(\frac{\mu R^2}{\varepsilon}\right)\right\rceil\right\}\right),$$

где, как и раньше,

$$R = \|x^0 - x_*\|_2, \quad \Delta f = f(x^0) - f(x_*).$$

Данная оценка в общем случае не может быть улучшена, даже если дополнительно известно, что $M_2 < \infty$, $M_3 < \infty$, … [74]. При этом данная оценка достигается [74, 76, 186, 477].

♦ Заметим, что если вместо $r \leqslant 1$ имеет место условие $r = 0$ (класс безградиентных методов), то в приведённой оценке все аргументы минимума следует домножить на размерность пространства n [9, 16, 32, 74, 88, 267, 268, 272, 415, 421, 487]. Отметим также, что у известных сейчас методов, отвечающих (с точностью до логарифмического множителя) первому аргументу минимума, достаточно дорогой является составляющая итерации, не связанная с вычислением градиента: $\gg n^2$ (см. указание к упражнению 1.4 и [74, 186, 419]). ♦

Для класса методов, у которых на каждой итерации разрешается не более чем O(1) раз обращаться к оракулу (подпрограмме) за значениями $\nabla^r f(x)$, $r \leqslant p$, $p \geqslant 2$, оценка числа итераций, необходимых для достижения точности ε (по функции), будет иметь вид [25][9] (в работе [192] приведено обобщение выписанной оценки на случай монотонных вариационных неравенств)

$$\mathrm{O}\left(\min\left\{n\ln\left(\frac{\Delta f}{\varepsilon}\right);\ \frac{M_0^2R^2}{\varepsilon^2},\ \left(\frac{M_1R^2}{\varepsilon}\right)^{1/2},\ \left(\frac{M_2R^3}{\varepsilon}\right)^{2/7},\ \ldots,\ \left(\frac{M_pR^{p+1}}{\varepsilon}\right)^{\frac{2}{3p+1}};\right.\right.$$

$$\min\left\{\left(\frac{M_1}{\mu}\right)^{1/2},\ \left(\frac{M_2R}{\mu}\right)^{2/7},\ \ldots,\ \left(\frac{M_pR^{p-1}}{\mu}\right)^{\frac{2}{3p+1}}\right\} +$$

$$\left.\left. + \min_{r=2,\ldots,p}\left\lceil\log\left\lceil\log\left(\frac{(\mu^{r+1}/M_r^2)^{1/(r-1)}}{\varepsilon}\right)\right\rceil\right\rceil\right\}\right). \quad (*)$$

[9] Примечательно, что этот результат был получен независимо и приблизительно в одно время (во второй половине 2018 года) разными коллективами (из МФТИ, Microsoft Research и коллективом авторов из Китая) под сильным влиянием работ Ю. Е. Нестерова и непосредственного общения с ним. Объединяющая публикация появилась в 2019 году на конференции COLT [291].

Данная оценка в общем случае не может быть улучшена, даже если дополнительно известно, что $M_{p+1} < \infty$, $M_{p+2} < \infty$, … [74, упражнение 2, § 1, гл. 8], [144, 392]. Полезно отметить, что здесь, так же как и для градиентных методов (см. упражнения 1.3, 2.1), можно строить «универсальные худшие в мире функции» в классе выпуклых полиномов (от абсолютных значений линейных комбинаций переменных) степени $p+1$ [144, 473]:

$$f_m(x) = \eta_{p+1}(A_m x) - x_1, \quad \eta_{p+1}(x) = \frac{1}{p+1}\sum_{i=1}^{n} |x_i|^{p+1},$$

$$A_m = \begin{pmatrix} U_m & 0 \\ 0 & I_{n-m} \end{pmatrix},$$

$$U_m = \underbrace{\begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & -1 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix}}_{m}, \quad I_{n-m} = \underbrace{\begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix}}_{n-m}.$$

В качестве точки старта выбирается $x^0 = (0, \dots, 0)^T$, а класс допустимых методов (алгоритмов) определяется условием

$$x^{k+1} \in x^0 + \sum_{l=0}^{k} S_f(x^l) \subseteq \{x \in \mathbb{R}^n : x_i = 0,\ i = k+2, \dots, n\},$$

где

$$S_f(x) = \operatorname{Lin}\left\{y_x(a_0, \dots, a_p; \gamma; q) : a_0, \dots, a_p \in \mathbb{R};\ \gamma > 0;\ q = 1, 2, 3, \dots\right\},$$

$$y_x(a_0, \dots, a_p; \gamma; q) \in \operatorname{Arg}\min_{y \in \mathbb{R}^n}\left\{\sum_{r=0}^{p} a_r \nabla^r f(x) \underbrace{[y-x, \dots, y-x]}_{r} + \gamma \|y-x\|_2^q\right\}.$$

В [473, теорема 4] было показано, что для всех методов из описанного выше класса при $N \leqslant (n-1)/2$ имеет место оценка

$$f_{2N+1}(x^N) - \min_{x \in \mathbb{R}^n} f_{2N+1}(x) \geqslant \frac{3p}{2^{p+1}(p+1)!} \frac{M_p R^{p+1}}{(2N+2)^{(3p+1)/2}},$$

где $M_p = p!$. При этом недавно было подмечено [144, 450], [477, п. 4.3], что данная оценка с точностью до числового множителя достигается при $p = 2$. В работах [150, 473] было показано, как при $p \geqslant 2$ получать методы (подобно ускоренному методу Ньютона с кубической регуляризацией [468]) с почти оптимальной оценкой скорости сходимости. Также в работах [314, 473, 475] было показано, что при $p = 3$ (случай $p = 2$ рассмотрен в работах [468, 475, 480]) трудоёмкость каждой ите-

рации предложенного метода сопоставима (с точностью до логарифмических множителей) с трудоёмкостью итерации метода Ньютона.

В работе [113] приводятся отличные от [144] нижние оценки при более общих условиях.

До недавнего времени считалось, что ускорение, напрямую унаследованное от градиентных методов, не даёт оптимальных оценок [150, 468, 473]. Тем не менее в 2018 г. задача построения оптимальных методов высокого порядка (работающих согласно ранее приведённой общей оценке) с помощью проксимального ускоренного метода Монтейро — Свайтера была решена, см. замечание 3.3 [25, 26, 189, 277, 307, 361, 477].

Поясним, как получается сильно выпуклая часть оценки $(*)$ (см. с. 219) из не сильно выпуклой[10]. Для этого рассмотрим *метод Ньютона*:

$$x^{k+1} = \arg\min_{x\in\mathbb{R}^n} \left\{ f(x^k) + \langle \nabla f(x^k), x - x^k\rangle + \frac{1}{2}\langle \nabla^2 f(x^k)(x-x^k), x-x^k\rangle \right\} = \\ = x^k - [\nabla^2 f(x^k)]^{-1} \nabla f(x^k).$$

♦ В методе Ньютона на каждой итерации предлагается минимизировать параболоид (уже, вообще говоря, не вращения), касающийся (до второго порядка включительно, т. е. в точке касания совпадают не только градиенты, но и гессианы) графика рассматриваемой функции. Однако этот параболоид, как правило, не мажорирует рассматриваемую функцию. Это обстоятельство не позволяет гарантировать глобальную сходимость метода. Одним из возможных решений проблемы глобальной сходимости методов второго порядка такого типа является добавление к параболоиду кубического слагаемого — кубической регуляризации [150, 468, 477, 480]. Полученная в результате вспомогательная задача (минимизации регуляризованного кубическим членом параболоида) по сложности сопоставима с исходной [200, 480], но при этом метод приобретает глобальную сходимость. Отметим, что локальная сходимость остаётся по-прежнему сверхлинейной, как у метода Ньютона.

Заметим, что на примере метода Ньютона для поиска корней простейших полиномов (вида z^3-1) можно продемонстрировать, насколько порой причудливыми бывают бассейны притяжения (бассейны/фракталы Ньютона) численных методов оптимизации в многоэкстремальных задачах [350]. ♦

[10] Не сильно выпуклая часть оценки (кроме первых двух аргументов минимума) получается из сильно выпуклой с помощью регуляризации $\mu \simeq \varepsilon/R^2$ (см. замечание 4.1).

Считая, что $M_2 < \infty$, $\mu > 0$, получим

$$\|\nabla f(x^{k+1})\|_2 = \left\|\nabla f(x^{k+1}) - \nabla f(x^k) - \nabla^2 f(x^k)(x^{k+1} - x^k)\right\|_2 \leqslant$$
$$\leqslant M_2 \|x^{k+1} - x^k\|_2^2 = M_2 \left\|[\nabla^2 f(x^k)]^{-1} \nabla f(x^k)\right\|_2^2 \leqslant \frac{M_2}{\mu^2} \|\nabla f(x^k)\|_2^2.$$

С помощью последнего неравенства можно оценить окрестность квадратичной скорости сходимости метода Ньютона:

$$\frac{M_2}{\mu^2}\|\nabla f(x^k)\|_2 < 1 \;\Rightarrow\; \frac{\mu}{2}\|x^k - x_*\|_2^2 \leqslant f(x^k) - f(x_*) \leqslant \frac{1}{2\mu}\|\nabla f(x^k)\|_2^2 < \frac{\mu^3}{2M_2^2}.$$

Значит, окрестность квадратичной скорости сходимости содержит открытый шар

$$\|x - x_*\|_2^2 < \frac{\mu^2}{M_2^2}.$$

Оказавшись в этой окрестности, можно достичь желаемой точности за

$$\mathrm{O}\left(\left\lceil \log \left\lceil \log \left(\frac{\mu^3}{M_2^2 \varepsilon} \right) \right\rceil \right\rceil\right)$$

итераций метода Ньютона.

♦ Заметим, что если последовательность чисел $\{c_k\}_{k=0,1,2,\ldots}$, $c_k > 0$, удовлетворяет условию (обеспечивающему *сверхлинейную скорость сходимости*)

$$c_{k+1} \leqslant \mathrm{const} \cdot (c_k)^\gamma, \quad \gamma > 1,$$

и c_0 достаточно мало, то после $N = \mathrm{O}(\log\lceil \log(c_0/\varepsilon)\rceil)$ итераций мы получаем $c_N \leqslant \varepsilon$. Для метода Ньютона $c_k = \|\nabla f(x^k)\|_2$, $\gamma = 2$; для класса чебышёвских методов высокого порядка $c_k = \|x^k - x_*\|_2$, $\gamma = 3, 4, 5, \ldots$ [49, п. 2.9], [59, п. 9.5.10], [244]; для методов Ньютона с кубической регуляризацией $c_k = f(x^k) - f(x_*)$, $\gamma = 4/3$, причём в последнем случае сильную выпуклость можно заменить градиентным доминированием [75, гл. 4], [468, 480] (см. также замечание 1.1). ♦

Чтобы оказаться в этой окрестности, можно использовать технику рестартов (см. упражнение 2.3, § 5, а также [25], [75, гл. 4], [144, 468]), применённую к методам, обеспечивающим несильно выпуклую составляющую рассматриваемой оценки.

Пусть

$$f(x^N) - f(x_*) \leqslant \frac{C M_p \|x^0 - x_*\|_2^{p+1}}{N^{(3p+1)/2}}.$$

В силу μ-сильной выпуклости функции $f(x)$ отсюда следует, что

$$\frac{\mu}{2}\|x^N - x_*\|_2^2 \leqslant f(x^N) - f(x_*) \leqslant \frac{C M_p \|x^0 - x_*\|_2^{p+1}}{N^{(3p+1)/2}}.$$

Выберем N таким образом, чтобы выполнялось неравенство

$$\|x^N - x_*\|_2^2 \leqslant \frac{1}{2}\|x^0 - x_*\|_2^2.$$

Для этого достаточно сделать

$$\left(\frac{4CM_pR^{p-1}}{\mu}\right)^{2/(3p+1)}$$

итераций. Процедуру можно повторить (рестартовать). Рестарты заканчиваются в момент попадания в окрестность квадратичной скорости сходимости для метода Ньютона. Таким образом, число рестартов можно оценить сверху следующим образом:

$$\left\lceil \log_2\left(\frac{M_2^2R^2}{\mu^2}\right)\right\rceil.$$

При этом общее число итераций до попадания в окрестность квадратичной сходимости будет составлять (учитываем, что после каждого рестарта R^2 уменьшается как минимум в два раза)

$$\mathrm{O}\left(\left(\frac{M_pR^{p-1}}{\mu}\right)^{2/(3p+1)}\right),$$

что и объясняет вторую заключительную (сильно выпуклую) часть рассматриваемой оценки.

Вместо метода Ньютона здесь можно было бы использовать методы более высокого порядка сходимости [48, 480]. При попадании в окрестность сверхлинейной (квадратичной, кубической и т. д.) скорости сходимости таких методов желаемая относительная точность по функции $\tilde{\varepsilon}$ будет достигнута после $\sim \log\log(\tilde{\varepsilon}^{-1})$ [48, 74, 144] дополнительных итераций (основания логарифмов зависят от порядка выбранного метода).

К сожалению, на данный момент неизвестно, как можно в общем случае использовать сильную выпуклость исходной задачи оптимизации, чтобы быстрее (по сравнению с выпуклым случаем) решать вспомогательную подзадачу (минимизации многочлена Тейлора) на каждой итерации рассматриваемых (тензорных) методов, см. указание к упражнению 3.10.

Отметим также, что имеются обнадеживающие результаты относительно падения чувствительности тензорных методов к точности задания старших производных, что может быть полезно для приложений (рандомизированных) тензорных методов к задачам минимизации суммы большого числа слагаемых [242, 440, 597].

Замечание 4. Если вместо сильной выпуклости и достаточной гладкости предполагать *самосогласованность* оптимизируемой функции, то, используя при анализе скорости сходимости вместо евклидовой нормы специальную локальную норму Дикина [47]

$$\|h\|_x = \langle h, \nabla^2 f(x) h\rangle^{1/2},$$

с помощью *демпфированного метода Ньютона* [76, п. 4.1.5]

$$x^{k+1} = x^k - \frac{1}{1 + \|\nabla f(x^k)\|_{x^k,*}} [\nabla^2 f(x^k)]^{-1} \nabla f(x^k)$$

можно получить такую оценку числа итераций [76, 164]:

$$\mathrm{O}\Big(f(x^0) - f(x_*) + \Big\lceil \log \Big\lceil \log\Big(\frac{1}{\varepsilon}\Big) \Big\rceil \Big\rceil \Big).$$

Более того, с помощью специального варианта метода отслеживания траектории [477, п. 5.2.2] можно улучшить последнюю оценку в части

$$\mathrm{O}(f(x^0) - f(x_*)) \to \widetilde{\mathrm{O}}\big(\sqrt{f(x^0) - f(x_*)}\big).$$

Введём $g(t) = f(w + t\mathrm{v})$. Самосогласованность функции $f(x)$ означает, что для любых w и v справедливо неравенство

$$|g'''(t)| \leqslant 2(g''(t))^{3/2}$$

при всех t, причём множитель 2 здесь выбран для определённости. Последнее неравенство эквивалентно неравенству

$$\big|\nabla^3 f(x)[h,h,h]\big| \leqslant 2\|h\|_x^3.$$

В частности, если для $f(x)$ выполняется условие $M_2 \leqslant 2\mu^{3/2}$, то функция $f(x)$ самосогласованная. В сильно выпуклом случае последнее неравенство можно обеспечить, например, шкалированием $f(x)$, учитывая, что левая и правая части неравенства будут по-разному реагировать на такое шкалирование. Заметим, что из ранее полученных оценок и простого наблюдения

$$\frac{\mu}{2} R^2 = \frac{\mu}{2}\|x^0 - x_*\|_2^2 \leqslant f(x^0) - f(x_*) = \Delta f,$$

т. е. $R^{p-1} \leqslant (2\Delta f/\mu)^{(p-1)/2}$, следует, что

$$\mathrm{O}\Big(\Big(\frac{M_p R^{p-1}}{\mu}\Big)^{2/(3p+1)}\Big) \to \mathrm{O}\Big(\Big(\frac{M_p \Delta f^{(p-1)/2}}{\mu^{(p+1)/2}}\Big)^{2/(3p+1)}\Big).$$

Таким образом, если $M_2 \leqslant 2\mu^{3/2}$, то при $p = 2$ мы имеем

$$\mathrm{O}(\Delta f^{1/7}) \ll \widetilde{\mathrm{O}}(\Delta f^{1/2}) \quad (\Delta f \gg 1),$$

что значительно лучше оценки сложности метода отслеживания траектории для задач оптимизации самосогласованных функций (см. также [274], [477, гл. 5]).

Отметим, что в описанном подходе константы имеют «физическую» размерность, поэтому, в отличие от формул, встречавшихся до настоящего момента, в этом замечании и оставшейся части пособия не стоит пытаться проверять корректность формул из соображений размерности [53, гл. 1]. Если дополнительно предполагать, что оптимизируемая функция является *ν-самосогласованным барьером*, т. е.

$$\nu\nabla^2 f(x) \succ \nabla f(x)\nabla f(x)^T,$$

то в классе методов 2-го порядка (см. ниже описание метода Нестерова — Немировского) также можно достичь следующей оценки числа итераций [76, 164, 477]:

$$O\left(\sqrt{\nu}\ln\left(\frac{\nu}{\varepsilon}\right)\right).$$

Для любого выпуклого замкнутого множества с непустой внутренностью, не содержащего прямых, можно построить (универсальный) самосогласованный барьер с $\nu \simeq n$ [76, теорема 4.3.2], [107, 187]. В подавляющем большинстве интересных для практики случаев удаётся конструктивно (явно) построить соответствующий самосогласованный барьер. ■

♦ В продолжение темы, затронутой в замечании 4, опишем, пожалуй, самый известный и хорошо проработанный способ глобализации сходимости методов второго порядка, см. замечание 1.3. Излагаемые далее результаты восходят к работам А. С. Немировского и Ю. Е. Нестерова конца 80-х годов XX века, посвящённым развитию *методов внутренней точки / методов внутренних штрафов* (*барьеров*) [76, 164, 477, 522]. Заметим, что метод (внешних) штрафов был описан в замечании 4.3.

Пусть нужно решить с точностью по функции ε следующую задачу:

$$\langle c, x\rangle \to \min_{x\in Q},$$

где множество Q уже не предполагается множеством простой структуры в смысле начала § 2. Однако предполагается, что для этого множества можно построить *ν-самосогласованный барьер*, т. е. построить такую достаточно гладкую строго выпуклую функцию $F(x)$, которая, в частности, стремится к бесконечности при приближении аргумента к точкам границы множества Q изнутри [76, 164, 477]. Далее исходная задача заменяется однопараметрическим семейством задач (*метод*

продолжения по параметру, см., например, [56, § 5.3])

$$t\langle c, x\rangle + F(x) \to \min_x,$$

где в конечном итоге нужно, чтобы выполнялось условие $t \to \infty$. В упомянутом цикле работ Ю. Е. Нестерова и А. С. Немировского был предложен следующий вариант метода Ньютона:

$$t^{k+1} = \Big(1 + \frac{1}{13\sqrt{\nu}}\Big)t^k, \quad x^{k+1} = x^k - [\nabla^2 F(x^k)]^{-1}\big(t^{k+1}\cdot c + \nabla F(x^k)\big),$$

который после $O(\sqrt{\nu}\ln\varepsilon^{-1})$ итераций находит с точностью по функции ε решение исходной задачи, если начальное приближение было разумно выбрано. Также было показано, что получить «разумное» начальное приближение всегда можно, сделав дополнительно не более $O(\sqrt{\nu}\ln\nu)$ итераций такого же типа, см., например, [186, п. 5.3].

Заметим, что исходную постановку задачи на самом деле можно завязать на более привычные по данному пособию постановки задачи. В частности, задачу

$$f(x) \to \min_x$$

можно эквивалентным образом переписать как

$$y \to \min_{f(x)\leqslant y}.$$

Описанный выше подход лежит в основе пакета выпуклой оптимизации CVX [621], о котором упоминалось во введении. Дело в том, что основные функции, возникающие в выпуклой оптимизации, можно единообразно записывать, используя следующее обобщение *представления* (*исключения*) *Фурье — Моцкина* [102, 164, 467], см. также замечание 3.1:

$$f(x) = \min_{y:\, A\left[\begin{smallmatrix}x\\y\end{smallmatrix}\right]=b,\ \left[\begin{smallmatrix}x\\y\end{smallmatrix}\right]\in K} \langle c, x\rangle + \langle d, y\rangle,$$

где выпуклый конус K есть прямое произведение конусов трёх (канонических) типов: $\mathbb{R}^n_+$, S^n_+, L^n_2, т. е. неотрицательного ортанта, конуса неотрицательно определённых матриц и лоренцевского конуса (ice cream cone). При этом, как было показано А. С. Немировским и А. Бен-Талем в 1998 г. [164, 165, 467], лоренцевский конус L^n_2 можно заменить конусом $\mathbb{R}^{O(n\ln\varepsilon^{-1})}_+$. Кроме того, L^n_2 можно понимать как прообраз конуса S^n_+ при специальном аффинном преобразовании, т. е. конус L^n_2 может быть полностью исключён из числа «канонических» (базисных) [164, п. 3.2]. Важно отметить, что для большого числа классов задач выпуклой оптимизации удалось найти такое представление.

Изначально был заготовлен достаточно большой набор стандартных (библиотечных) выпуклых функций с известными, найденными «вручную» представлениями. Затем были определены правила сочетания, которые позволяют по известным функциям получать новые, и написана программа, которая организует разумный перебор этих правил для поиска представления для новых функций, не присутствующих в библиотеке. Детали см., например, в [313, 467].

Отметим, что самосогласованность, вообще говоря, не подразумевает сильную выпуклость. В частности, для задачи квадратичной оптимизации (возникающей, например, при приближённом представлении функции $f(x) = e^x$ с помощью лоренцевского конуса [467])

$$x_n \to \min_{\substack{x_1 \geqslant 0, x_2 \geqslant x_1^2, \\ x_3 \geqslant x_2^2, \ldots, x_n \geqslant x_{n-1}^2}}$$

можно явно построить самосогласованный барьер и наблюдать на практике хорошее соответствие с теорией в части сходимости описанного метода по функции. Однако сходимость по аргументу чрезвычайно замедляется с ростом размерности пространства n, в котором происходит оптимизация. Схожие проблемы и дополнительная вычислительная неустойчивость возникают и при работе с конусом неотрицательных полиномов [489], играющим важную роль в теории управлении [327] и возникающим также при разработке оптимальных алгоритмов [541]. Особенностью данного конуса является наличие двойственного представления (конуса), порождённого специальным подклассом неотрицательно определённых ганкелевых матриц. Этот двойственный конус практически не пересекает конус неотрицательно определённых матриц, и работа с ним приводит к серьёзным вычислительным трудностям. В качестве простого примера укажем, что популярная в задачах анализа данных плохо обусловленная матрица Гильберта [86, п. 5, § 1, гл. 11] является одним из самых безобидных представителей данного класса ганкелевых матриц [589].

Таким образом, удалось автоматически редуцировать подавляющее большинство задач выпуклой оптимизации к единому виду:

$$\langle c, x \rangle + \langle d, y \rangle \to \min_{(x,y) : A\begin{bmatrix} x \\ y \end{bmatrix} = b,\ \begin{bmatrix} x \\ y \end{bmatrix} \in K}.$$

Чтобы применять вариант метода внутренней точки, описанный выше, осталось только подобрать самосогласованные барьеры для канонических конусов $\mathbb{R}^n_+$, S^n_+ и понять, что происходит при их линейных преобразованиях. Но эти задачи уже были успешно решены ещё

в самых первых работах. В частности, были построены следующие n-самосогласованные барьеры:

$$F_{\mathbb{R}^n_+}(x) = -\sum_{i=1}^{n} \ln x_i, \quad F_{S^n_+}(X) = -\ln \det(X).$$

Таким образом и появился один из самых популярных сейчас пакетов CVX решения задач выпуклой оптимизации умеренных размеров — до десятков тысяч переменных [621]. На наш взгляд, оболочка CVX имеет наилучшую реализацию для использования вместе с MatLab и не очень удачную реализацию для работы с Python. Отметим также другие оболочки и солверы, которые на протяжении многих лет являются помощниками специалистов в области численных методов оптимизации. Среди оболочек выделим также AMPL [624] и GAMS [635], а среди солверов отметим CPLEX [636], MOSEK [638], GUROBI Optimization [623], Local Solver [637]. Имеется большой сервер солверов [630], на котором на данный момент можно найти более 60 различных оптимизационных солверов, в том числе и отмеченные выше. И конечно, нельзя не упомянуть открытую программную библиотеку Google, заточенную под задачи машинного обучения, в частности задачи глубокого обучения: TensorFlow [639] (см. также [627, 629, 631]).

Заинтересовавшемуся в замечании 4 и последующем тексте читателю можно рекомендовать ознакомиться с книгой [76, гл. 4] и работами [164, п. 2.3, 2.5, 3.2], [186, п. 5.3] для более глубокого погружения в затронутые здесь темы. Особо отметим то, насколько неожиданными порой бывают выпуклые функции, допускающие явно выписываемое обобщённое представление Фурье — Моцкина [467]. ♦

Так же как и для методов 1-го порядка (см. § 5), для методов 2-го порядка и выше можно рассматривать их универсальные композитные варианты [317, 318], а также рассматривать работу методов в условиях наличия шума и неточностей, возникающих при решении вспомогательных задач на каждой итерации [150, 297]; кроме того, можно переносить на них и завязанные на наличие шумов конструкции, например конструкцию *минибатчинга* [297].

Однако при использовании методов 2-го порядка и выше появляется много новых вопросов относительно сильного проигрыша методам первого порядка (градиентного типа) по стоимости итерации и требуемой памяти. Так, для честного осуществления шага метода Ньютона необходимо обратить матрицу Гессе оптимизируемой функции в текущей точке. Эта задача по сложности эквивалентна задаче

умножения двух матриц такого же по порядку размера [63, гл. 31], что в типичном случае в n раз дороже, чем осуществление шага метода типа градиентного спуска (умножение матрицы на вектор).

♦ На самом деле это не так. Умножение двух матриц размера $n \times n$ современными алгоритмами может быть осуществлено за время $O(n^{2,37})$, см. [89, 288, 380] и цитированную там литературу. Однако такого рода результаты проявляются только при очень больших значениях n. ♦

В последнее время было предложено несколько подходов, имеющих своей целью хотя бы частичное устранение такого большого зазора в стоимости итерации между методами 1-го и 2-го порядка. Одна из идей активно используется в машинном обучении, когда функционал имеет вид суммы (среднего арифметического) большого числа однотипных слагаемых. Идея заключается в том, чтобы формировать матрицу Гессе оптимизируемой функции исходя из матриц Гессе относительно небольшого числа случайно выбранных слагаемых [297, 597]. Другая идея заключается в отказе от обращения матрицы Гессе на итерации, и вместо этого предлагается использовать информацию о собственном векторе, отвечающем наименьшему собственному значению [112, 196]. Для приближённого вычисления такого вектора вполне достаточно уметь умножать матрицу Гессе на произвольный вектор [125]:

$$\nabla^2 f(x)\mathrm{v} = \nabla \langle \nabla f(x), \mathrm{v} \rangle \approx \frac{\nabla f(x + \tau \mathrm{v}) - \nabla f(x)}{\tau},$$

что может быть сделано с помощью автоматического дифференцирования за то же по порядку время, что и вычисление градиента [159, 495]. Эта идея сейчас активно развивается в связи с поиском наиболее эффективных методов обучения глубоких нейронных сетей [40, 82, 125, 126, 196].

♦ Одной из основных проблем, возникающих при обучении глубоких нейронных сетей, является «застревание» процедур типа градиентного спуска в «паразитных» экстремумах [40] (седловых точках). В качестве примера, следуя работе [608], приведём способ «проскальзывания седловых точек» с помощью доступа к $\nabla^2 f(x)\mathrm{v}$. Будем считать, что $M_1, M_2 < \infty$, см. выше. Следуя работе [480], будем искать такую точку x^N ((ε, δ)-локальный минимум), что

$$\|\nabla f(x^N)\|_2 \leqslant \varepsilon, \quad \lambda_{\min}(\nabla^2 f(x^N)) \geqslant -\delta.$$

Рассмотрим следующую процедуру:

$$x^{k+1} = x^k - \frac{1}{M_1} \nabla f(x^k), \quad \text{если} \quad \|\nabla f(x^k)\|_2 > \varepsilon;$$

$$x^{k+1} = x^k + hp^k, \quad \text{если} \quad \|\nabla f(x^k)\|_2 \leqslant \varepsilon \ \text{ и } \ \lambda_{\min}(\nabla^2 f(x^k)) < -\delta,$$

где $h = 2\delta/M_2$, p^k — собственный вектор для $\nabla^2 f(x^k)$, отвечающий наименьшему собственному значению. Вектор p^k определяется с точностью до произвола в нормировке. Распорядимся этим произволом следующим образом:

$$\langle \nabla f(x^k), p^k \rangle \leqslant 0, \quad \|p^k\|_2 = 1.$$

Из определения M_2 следует (см., например, [473]), что если «сработала» вторая альтернатива, то

$$f(x^{k+1}) \leqslant f(x^k) + h \underbrace{\langle \nabla f(x^k), p^k \rangle}_{\leqslant 0} + \frac{h^2}{2!} \underbrace{\langle \nabla^2 f(x^k) p^k, p^k \rangle}_{\leqslant -\frac{h^2\delta}{2}} + \frac{h^3}{3!} \underbrace{M_2 \|p^k\|_2^3}_{=M_2} \leqslant$$

$$\leqslant f(x^k) - \frac{1}{2}\left(\frac{2\delta}{M_2}\right)^2 \delta + \frac{M_2}{6}\left(\frac{2\delta}{M_2}\right)^3 = f(x^k) - \frac{2}{3}\frac{\delta^3}{M_2^2}.$$

Повторяя далее рассуждения из §1, которые использовались при выводе формулы (1.10), получим, что метод гарантированно остановится после

$$N = \left(f(x^0) - f(x^{\text{local min}})\right) \cdot \max\left\{\frac{2M_1}{\varepsilon^2}, \frac{3M_2^2}{2\delta^3}\right\}$$

итераций. Заметим, что искать в данной процедуре собственный вектор точно нет необходимости. Достаточно либо определить, что

$$\lambda_{\min}(\nabla^2 f(x^k)) \geqslant -\delta,$$

либо найти такой вектор p^k ($\|p^k\|_2 = 1$), что $\langle \nabla^2 f(x^k) p^k, p^k \rangle \leqslant -\delta/2$. Эта задача может быть решена обычным степенным методом (методом простой итерации), тесно связанным с методом градиентного спуска (см. замечание 4.4), за $O(1/\delta)$ итераций (на каждой итерации происходит умножение $\nabla^2 f(x)v$) и методом Ланцоша, тесно связанным с методом сопряжённых градиентов [95, п. 21.4], за $O(1/\sqrt{\delta})$ аналогичных итераций.

Совсем грубо это можно пояснить, например, так. Пусть известно (в этом месте и делается грубое допущение, поскольку знать $\lambda_{\min}$ на самом деле не нужно) $\lambda_{\min} = \lambda_{\min}(\nabla^2 f(x^k)) < -\delta$. Рассмотрим вспомогательную задачу

$$\frac{1}{2}\langle \nabla^2 f(x^k) p, p \rangle + \frac{|\lambda_{\min}|}{2}\|p\|_2^2 \to \min_p.$$

Подобно замечанию 4.4 можно показать, что градиентные методы (в том числе ускоренные) сходятся для данной задачи (вырожденной,

но выпуклой!) к собственному вектору, отвечающему $\lambda_{\min}$, наиболее близкому (в 2-норме) к точке старта, которую в этой связи следует выбирать отличной от нуля. Осталось только заметить следующее: для поиска такого вектора p^k ($\|p^k\|_2 = 1$), что

$$\langle \nabla^2 f(x^k) p^k, p^k \rangle \leqslant -\frac{\delta}{2} \cdot \|p^k\|_2^2,$$

достаточно решить ускоренным методом вспомогательную задачу с точностью (по функции) $(\delta/4) \cdot \|p^k\|_2^2$. Отсюда и получается оценка $O(1/\sqrt{\delta})$. Отметим цикл работ [184, 556, 557], в которых приводятся нижние оценки для класса *методов Крылова* [95, 211], использующих только умножение матрицы на вектор, для решения систем линейных уравнений и задачи вычисления собственного вектора матрицы, отвечающего максимальному (минимальному) собственному значению, эквивалентной ей по сложности с точностью до логарифмического множителя. Нижние оценки получаются в полном соответствии с оценками, полученными отмеченной выше редукцией исходной постановки задачи к задаче квадратичной выпуклой оптимизации.

Полученную выше оценку на число итераций $O(1/\sqrt{\delta})$ можно улучшить, в особенности если целевой функционал имеет вид суммы большого числа слагаемых. В этом случае используются различные инкрементальные (рандомизированные) методы, см. замечание 1 и работы [122, 125, 126, 277, 619], а также цитированную в этих работах литературу.

Обычно различают две постановки задачи.

1. В предположении, что везде $\lambda_{\min}(\nabla^2 f(x)) \geqslant -\sigma$, где $\sigma > 0$ — малая величина, требуется найти такую точку x^N, что $\|\nabla f(x^N)\|_2 \leqslant \varepsilon$. (Задача поиска глобального ε-минимума по функции даже в такой постановке остаётся NP-трудной [456].)
2. В условиях достаточной гладкости целевой функции (слагаемых) требуется найти (ε, δ)-локальный минимум (см. выше).

По постановке 1 уже построена более-менее законченная теория в классе оффлайн-методов[11]: получены нижние оценки и предложены методы, работающие по нижним оценкам с точностью до логарифмических множителей, см. работу [618] и цитированную там литературу.

[11] Использована терминология работ [122, 126, 277]. Скорость сходимости оффлайн-методов может зависеть от числа слагаемых, см. упражнение 1.8. При этом можно ограничиться классом методов для задач минимизации функционала вида суммы, сложность которых не зависит от числа слагаемых, см., например, замечание 1 и упражнение 1.8.

Отметим используемую в ряде подходов (см., например, работу [122] и цитированную там литературу) специальную технику итеративной регуляризации, напоминающую проксимальный подход (см. замечания 3.2, 3.3):

$$x^{k+1} \approx \arg\min_x \left\{ f(x) + \sigma \|x - x^k\|_2^2 \right\},$$

где вспомогательная задача уже будет $\sigma/2$-сильно выпуклой, следовательно, её можно решать ускоренными методами [412, 432].

По постановке 2 в последние годы активно ведутся исследования [122, 125, 126, 128, 140]. Наилучшие известные сейчас методы приведены в работах [140, 277, 619] и упражнении 1.8 (при должной доработке [277]), см. также работы [210, 509, 598, 599], содержащие практически более эффективные варианты.

Отметим интересное обобщение упомянутого ранее метода Ланцоша на случай, когда вместо произведения $\nabla^2 f(x)\mathrm{v}$ доступно произведение $\nabla^2 f_i(x)\mathrm{v}$ для случайно (равновероятно) выбранного слагаемого. В работе [125] была предложена разновидность алгоритма Ойя (рандомизированный вариант метода Ланцоша), позволяющая решать ту же задачу, что решалась выше с помощью метода Ланцоша, за $\widetilde{\mathrm{O}}(1/\delta^2)$ операций произведения вида $\nabla^2 f_i(x)\mathrm{v}$. При этом такие произведения не обязательно специально генерировать, их можно «собирать» по ходу работы основного алгоритма, как бы «заодно».

Отметим, что развиваемые в этом направлении наивные варианты методов типа редукции дисперсии в случае существенно невыпуклой целевой функции могут вести себя очень плохо на практике [229]. ♦

Перспективной также представляется довольно старая идея глобализации сходимости [46, 56, 217, 495]: спуск в область квадратичной (в общем случае сверхлинейно) сходимости с помощью методов типа градиентного спуска (с дешёвыми итерациями) и последующая квадратичная сходимость с использованием, например, метода Ньютона. Проблема в таком подходе — детектирование момента попадания в нужную окрестность. В качестве возможного решения проблемы можно, например, действовать таким образом: через каждые $\sim \sqrt{n}$ итераций метода типа градиентного спуска проверять условие

$$\|\nabla f(x^k)\|_2 \ll 1,$$

а если оно выполняется, то делать «пристрелочный» шаг метода Ньютона. Если в результате такого шага выполняется ещё и условие

$$\|\nabla f(x^{k+1})\|_2 \ll \|\nabla f(x^k)\|_2^{3/2},$$

то продолжать делать шаги метода Ньютона, каждый раз проверяя это условие. Если хотя бы одно из этих условий не выполняется, то следует вернуться к методу типа градиентного спуска.

На этом сюжете заканчивается изложение. В заключение отметим, что современные численные методы оптимизации являются бурно развивающейся областью исследований. В 2010 году в издательстве МЦНМО вышла книга (учебник по численным методам выпуклой оптимизации) Ю. Е. Нестерова [76]. Книга стала настольной для многих специалистов по вычислительной математике и анализу данных. Однако за прошедшие с момента написания книги более 10 лет было получено много новых важных результатов (см., например, `https://blogs.princeton.edu/imabandit/2019/12/30/a-decade-of-fun-and-learning/`), которые прочно вошли в основной арсенал современных специалистов по численным методам оптимизации. Настоящее приложение и многочисленные замечания, разбросанные по всей книге, имели одной из своих целей дать некоторое представление об этих новых результатах. Вряд ли стоит переоценивать роль этих «заметок». Знакомство с ними, скорее всего, не даст сразу полного понимания стоящих за этими заметками сюжетов. Однако наличие таких «отступлений» может упростить и систематизировать дальнейшее, более предметное, изучение современных численных методов оптимизации. В частности, таких популярных сейчас направлений как стохастическая оптимизация, распределённая оптимизация и невыпуклая оптимизация.

Надеемся, что данное пособие вызовет желание поработать в описанных в нём направлениях!

Литература

1. *Агаев Р. П.*, *Чеботарёв П. Ю.* Сходимость и устойчивость в задачах согласования характеристик (обзор базовых результатов) // Управление большими системами. 2010. Т. 30, № 1. С. 470–505.
2. *Алкуса М. и др.* Ускоренные методы для седловых задач // ЖВМ и МФ. 2020. Т. 60, № 11. С. 1787–1809.
3. *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.
4. *Аникин А. С. и др.* Двойственные подходы к задачам минимизации сильно выпуклых функционалов простой структуры при аффинных ограничениях // ЖВМ и МФ. 2017. Т. 57, № 8. С. 1270–1284.
5. *Антипин А. С.* Минимизация выпуклых функций на выпуклых множествах с помощью дифференциальных уравнений // Дифференциальные уравнения. 1994. Т. 30, № 9. С. 1395–1375.
6. *Бадриев И. Б.*, *Задворнов О. А.* Итерационные методы решения вариационных неравенств в гильбертовых пространствах. Казань: Казанский государственный университет, 2007.
7. *Баймурзина Д. Р. и др.* Универсальный метод поиска равновесий и стохастических равновесий в транспортных сетях // ЖВМ и МФ. 2019. Т. 59, № 1. С. 21–36.
8. *Бакушинский А. Б.*, *Кокурин М. Ю.* Итерационные методы решения нерегулярных уравнений. Учебное пособие по курсу «Математические методы системного анализа». М.: ЛЕНАНД, 2006.
9. *Баяндина А. С.*, *Гасников А. В.*, *Лагуновская А. А.* Безградиентные двухточечные методы решения задач стохастической негладкой выпуклой оптимизации при наличии малых шумов не случайной природы // Автоматика и Телемеханика. 2018. № 8. С. 38–49.
10. *Бирюков А. Г.* Методы оптимизации. Условия оптимальности в экстремальных задачах. М.: МФТИ, 2010.
11. *Бирюков С. И.* Оптимизация. Элементы теории и численные методы. М.: МЗ-Пресс, 2003.
12. *Брэгман Л. М.* Релаксационный метод нахождения общей точки выпуклых множеств и его применение для решения задач выпуклого программирования // ЖВМ и МФ. 1967. Т. 7, № 3. С. 200–217.
13. *Васильев Ф. П.* Методы оптимизации. Т. 1. М.: МЦНМО, 2011.
14. *Васильев Ф. П.* Методы оптимизации. Т. 2. М.: МЦНМО, 2011.

15. *Вершик А. М.*, *Спорышев П. В.* Оценки среднего числа шагов симплекс-метода и задачи асимптотической интегральной геометрии // Доклады Академии наук СССР. 1983. V. 271, № 5. С. 1044–1048.
16. *Воронцова Е. А.*, *Гасников А. В.*, *Горбунов Э. А.* Ускоренные спуски по случайному направлению с неевклидовой прокс-структурой // Автоматика и Телемеханика. 2019. № 4. С. 126–143.
17. *Воронцова Е. А. и др.* Поиск равновесия по Вальрасу и централизованная распределённая оптимизация с точки зрения современных численных методов выпуклой оптимизации на примере задачи распределения ресурсов // Сиб. журн. вычисл. матем. 2018. Т. 22, № 4. С. 415–436.
18. *Гасников А. В.*, *Гасникова Е. В.*, *Нестеров Ю. Е.* Двойственные методы поиска равновесий в смешанных моделях распределения потоков в больших транспортных сетях // ЖВМ и МФ. 2018. Т. 58, № 9. С. 1447–1454.
19. *Гасников А. В.*, *Двуреченский П. Е.*, *Камзолов Д. И.* Градиентные и прямые методы с неточным оракулом для задач стохастической оптимизации // Динамика систем и процессы управления. Труды Международной конференции, посвященной 90-летию со дня рождения академика Н. Н. Красовского. Екатеринбург, Россия. 15–20 сентября 2014. Институт математики и механики УрО РАН им. Н. Н. Красовского (Екатеринбург), 2015. С. 111–117.
20. *Гасников А. В.*, *Двуреченский П. Е.*, *Нестеров Ю. Е.* Стохастические градиентные методы с неточным оракулом // Труды МФТИ. 2016. Т. 8, № 1. С. 41–91.
21. *Гасников А. В.*, *Двуреченский П. Е.*, *Усманова И. Н.* О нетривиальности быстрых (ускоренных) рандомизированных методов // Труды МФТИ. 2016. Т. 8, № 2. С. 67–100.
22. *Гасников А. В. и др.* Адаптивный проксимальный метод для вариационных неравенств // ЖВМ и МФ. 2019. Т. 59, № 5. С. 889–894.
23. *Гасников А. В. и др.* Вокруг степенного закона распределения компонент вектора PageRank // Сиб. журн. вычисл. матем. 2017–2018. arXiv: 1701.02595.
24. *Гасников А. В. и др.* Об эффективных численных методах решения задач энтропийно-линейного программирования // ЖВМ и МФ. 2016. Т. 56, № 4. С. 523–534.
25. *Гасников А. В. и др.* Обоснование гипотезы об оптимальных оценках скорости сходимости численных методов выпуклой оптимизации высоких порядков // Компьютерные исследования и моделирование. 2018. Т. 10, № 6. С. 737–754.
26. *Гасников А. В. и др.* Ускоренный метаалгоритм для задач выпуклой оптимизации // ЖВМ и МФ. 2021. Т. 61, № 1 (в печати).
27. *Гасников А. В.*, *Камзолов Д. И.*, *Мендель М. А.* Основные конструкции над алгоритмами выпуклой оптимизации и их приложения к получе-

нию новых оценок для сильно выпуклых задач // Труды МФТИ. 2016. Т. 8, № 3. С. 25–42.

28. *Гасников А. В.*, *Кубентаева М. Б.* Поиск стохастических равновесий в транспортных сетях с помощью универсального прямо-двойственного градиентного метода // Компьютерные исследования и моделирование. 2018. Т. 10:3. С. 335–345.
29. *Гасников А. В.*, *Лагуновская А. А.*, *Морозова Л. Э.* О связи имитационной логит динамики в популяционной теории игр и метода зеркального спуска в онлайн оптимизации на примере задачи выбора кратчайшего маршрута // Труды МФТИ. 2015. Т. 7, № 4. С. 104–113.
30. *Гасников А. В.*, *Нестеров Ю. Е.*, *Спокойный В. Г.* Об эффективности одного метода рандомизации зеркального спуска в задачах онлайн оптимизации // ЖВМ и МФ. 2015. Т. 55, № 4. С. 582–598.
31. *Гасников А. В.*, *Нестеров Ю. Е.* Универсальный метод для задач стохастической композитной оптимизации // ЖВМ и МФ. 2018. Т. 58, № 1. С. 51–68.
32. *Гасников А. В.* Эффективные численные методы поиска равновесий в больших транспортных сетях: дис. … д-ра физ.-мат. наук: 05.13.18. М.: МФТИ, 2016.
33. *Гельфанд И. М.*, *Цетлин М. Л.* О некоторых способах управления сложными системами // Успехи мат. наук. 1962. Т. 17, № 1(103). С. 3–25.
34. *Гилл Ф.*, *Мюррей У.*, *Райт М.* Практическая оптимизация. М.: Мир, 1985.
35. *Гловински Р.*, *Лионс Ж. Л.*, *Тремольер Р.* Численное исследование вариационных неравенств. М.: Мир, 1979.
36. *Голиков А. И.*, *Евтушенко Ю. Г.*, *Моллаверди Н.* Применение метода Ньютона к решению задач линейного программирования большой размерности // ЖВМ и МФ. 2004. Т. 44, № 9. С. 1564–1573.
37. *Голуб Дж.*, *Ван Лоун Ч.* Матричные вычисления. М.: Мир, 1999.
38. *Горнов А. Ю.* Вычислительные технологии решения задач оптимального управления. Новосибирск: Наука, 2009.
39. *Граничин О. Н.*, *Поляк Б. Т.* Рандомизированные алгоритмы оценивания и оптимизации при почти произвольных помехах. М.: Наука, 2003.
40. *Гудфеллоу Я.*, *Бенджио И.*, *Курвиль А.* Глубокое обучение. ДМК Пресс, 2017.
41. *Данскин Дж. М.* Теория максмина. М.: Советское радио, 1970.
42. *Двинских Д. М. и др.* Ускоренный градиентный слайдинг в задачах минимизации суммы функций // Доклады Академии наук. Сер. математическая. 2020. Т. 492. С. 85–88.
43. *Двинских Д. М. и др.* Ускоренный и неускореный стохастический градиентный спуск в модельной общности // Матем. заметки. 2020. Т. 108, вып. 64. С. 515–528.
44. *Двуреченский П. Е.*, *Гасников А. В.*, *Лагуновская А. А.* Параллельные алгоритмы и оценки вероятностей больших уклонений в задачах стоха-

стической выпуклой оптимизации // Сиб. журн. вычисл. матем. 2018. Т. 21, № 1. С. 47–53.

45. *Демьянов В. Ф., Малоземов В. Н.* Введение в минимакс. М.: Наука, 1972.
46. *Денис Дж., Шнабель Р.* Численные методы безусловной оптимизации и решения нелинейных уравнений. М.: Мир, 1988.
47. *Дикин И. И.* Метод внутренних точек в линейном и нелинейном программировании. М.: КРАСАНД, 2010.
48. *Евтушенко Ю. Г.* Методы решения экстремальных задач и их применение в системах оптимизации. М.: Наука, 1982.
49. *Евтушенко Ю. Г.* Оптимизация и быстрое автоматическое дифференцирование. М.: ВЦ РАН, 2013.
URL: `http://www.ccas.ru/personal/evtush/p/198.pdf`
50. *Ермаков С. М.* Метод Монте-Карло в вычислительной математики. Вводный курс. СПб.: Невский диалект, 2009.
51. *Ермольев Ю. М.* Методы стохастического программирования. М.: Наука, 1976.
52. *Жадан В. Г.* Методы оптимизации. Ч. 1–3. М.: МФТИ, 2015–2017.
53. *Зорич В. А.* Математический анализ задач естествознания. М.: МЦНМО, 2017.
54. *Иванов В. К., Мельникова И. В., Филинков А. И.* Дифференциально-операторные уравнения и некорректные задачи. М.: Наука, 1995.
55. *Иванова А. С. и др.* Численные методы для задачи распределения ресурсов в компьютерной сети // ЖВМ и МФ. 2021. Т. 61, № 2.
56. *Измаилов А. Ф., Солодов М. В.* Численные методы оптимизации. М.: Физматлит, 2005.
57. *Измаилов А. Ф., Третьяков А. А.* Фактор-анализ нелинейных отображений. М.: Наука, 1994.
58. *Канторович Л. В., Акилов Г. П.* Функциональный анализ. СПб.: Невский диалект, 2004.
59. *Карманов В. Г.* Математическое программирование. М.: Наука, 1986.
60. *Ким К. и др.* Эффективные алгоритмы для дифференцирования и задачи экстремали // Экономика и математические методы. 1984. Т. 20. С. 309–318.
61. *Кириллов А. А., Гвишиани А. Д.* Теоремы и задачи функционального анализа. М.: Наука, 1988.
62. *Коннов И. В.* Нелинейная оптимизация и вариационные неравенства. Казань: Казанский ун-т, 2013.
63. *Кормен Т., Лейзерсон Ч., Ривест Р.* Алгоритмы: построение и анализ. М.: МЦНМО, 2002.
64. *Корпелевич Г. М.* Экстраградиентный метод для отыскания седловых точек и других задач // Экономика и мат. методы. 1976. Т. 12, № 4. С. 747–756.

65. *Лаврентьев М. М.*, *Романов В. Г.*, *Шишатский С. П.* Некорректные задачи математической физики и анализа. М.: Наука, 1980.
66. *Магарил-Ильяев Г. Г.*, *Тихомиров В. М.* Выпуклый анализ и его приложения. М.: УРСС, 2011.
67. *Матиясевич Ю. В.* Быстрая арифметика // Математическая составляющая / Редакторы-составители Н. Н. Андреев, С. П. Коновалов, Н. М. Панюнин; Художник-оформитель Р. А. Кокшаров. М.: Фонд «Математические этюды», 2015. URL: http://book.etudes.ru/toc/fast-arithmetic/
68. *Моисеев Н. Н.*, *Иванилов Ю. П.*, *Столярова Е. М.* Методы оптимизации. М.: Наука, 1978.
69. *Моисеев Н. Н.* Численные методы в теории оптимальных систем. М.: Наука, 1971.
70. *Назин А. В. и др.* Алгоритмы робастной стохастической оптимизации на основе метода зеркального спуска // Автоматика и Телемеханика. 2019. № 9. С. 64–90.
71. *Немировский А. С.*, *Нестеров Ю. Е.* Оптимальные методы гладкой выпуклой оптимизации // ЖВМ и МФ. 1985. Т. 25, № 3. С. 356–369.
72. *Немировский А. С.* О регуляризующих свойствах метода сопряжённых градиентов на некорректных задачах // ЖВМ и МФ. 1986. Т. 26, № 3. С. 332–347.
73. *Немировский А. С.*, *Поляк Б. Т.* Итерационные методы решения линейных некорректных задач при точной информации. II // Изв. АН СССР. Техническая кибернетика. 1984. № 3. С. 18–25.
74. *Немировский А. С.*, *Юдин Д. Б.* Сложность задач и эффективность методов оптимизации. М.: Наука, 1979.
75. *Нестеров Ю. Е.* Алгоритмическая выпуклая оптимизация: дис. … д-ра физ.-мат. наук: 01.01.07. М.: МФТИ, 2013.
76. *Нестеров Ю. Е.* Введение в выпуклую оптимизацию. М.: МЦНМО, 2010.
77. *Нестеров Ю. Е.* Метод минимизации выпуклых функций со скоростью сходимости χ // ДАН АН СССР. 1983. Т. 269, № 3. С. 543–547.
78. *Нестеров Ю. Е.* Нижние оценки сложности и оптимальные алгоритмы. Из курса лекций «Алгоритмические основы современной теории оптимизации». 2012. URL: http://www.mathnet.ru/php/seminars.phtml?option_lang=rus&presentid=6986
79. *Нестеров Ю. Е.*, *Скоков В. А.* К вопросу о тестировании алгоритмов безусловной оптимизации // Численные методы математического программирования. М.: ЦЭМИ, 1980. С. 77–91.
80. *Нестеров Ю. Е.* Эффективные методы в нелинейном программировании. М.: Радио и связь, 1989.
81. *Никайдо Х.* Выпуклые структуры и математическая экономика. М.: Мир, 1972.
82. *Николенко С. И.*, *Кадурин А. А.*, *Архангельская Е. О.* Глубокое обучение. Погружение в мир нейронных сетей. СПб.: Питер, 2018.

83. *Новиков П. С.* Элементы математической логики. М.: Наука, 1973.
84. *Нурминский Е. А.* Численные методы решения детерминированных и стохастических минимаксных задач. Киев: Наукова думка, 1979.
85. *Обен Ж.-П.* Нелинейный анализ и его экономические приложения. М.: Мир, 1988.
86. *Поляк Б. Т.* Введение в оптимизацию. М.: Наука, 1983.
87. *Поляк Б. Т.* Градиентные методы минимизации функционалов, решения уравнений и неравенств: дис. … канд. физ.-мат. наук. М.: МГУ, 1963.
88. *Протасов В. Ю.* К вопросу об алгоритмах приближенного вычисления минимума выпуклой функции по ее значениям // Мат. заметки. 1996. Т. 59, № 1. С. 95–102.
89. *Разборов А. А.* Алгебраическая сложность. М.: МЦНМО, 2016.
90. *Стецюк П. И.* Методы эллипсоидов и r-алгоритмы. Кишинэу: Эврика. 2014.
91. *Стонякин Ф. С.* Адаптивный аналог метода Ю. Е. Нестерова для вариационных неравенств с сильно монотонным полем // Сиб. журн. вычисл. матем. 2019. Т. 22, № 2. С. 201–211.
92. *Сухарев А. Г.*, *Тимохов А. В.*, *Федоров В. В.* Курс методов оптимизации. М.: Физматлит, 2005.
93. *Тер-Крикоров А. М.*, *Шабунин М. И.* Курс математического анализа. М.: БИНОМ. Лаборатория знаний, 2013.
94. *Тихонов А. Н.*, *Арсенин В. Я.* Методы решения некорректных задач. М.: Наука, 1979.
95. *Тыртышников Е. Е.* Методы численного анализа. М.: МГУ, 2006.
96. *Тюрин А. И.* Адаптивный быстрый градиентный метод в задачах стохастической оптимизации // arXiv:1712.00062.
97. *Тюрин А. И.*, *Гасников А. В.* Быстрый градиентный спуск для задач выпуклой минимизации с оракулом, выдающим (δ, L)-модель функции в запрошенной точке // ЖВМ и МФ. Т. 59, № 7. С. 1137–1150.
98. *Тюрин А. И.* Зеркальный вариант метода подобных треугольников для задач условной оптимизации // arXiv:1705.09809.
99. *Тюрин А. И.* Прямо-двойственный быстрый градиентный метод с моделью // arXiv:1906.10107.
100. *Уайлд Д. Дж.* Методы поиска экстремума. М.: Наука, 1967.
101. *Хачиян Л. Г.* Избранные труды / Сост. С. П. Тарасов. М.: МЦНМО, 2009.
102. *Циглер Г. М.* Теория многогранников. М.: МЦНМО, 2014.
103. *Червоненкис А. Я.* Компьютерный анализ данных. М.: Лекции Школы анализа данных Яндекс, 2009.
104. *Шор Н. З.* Методы минимизации недифференцируемых функции и их приложения. Киев: Наукова думка, 1979.
105. *Эванс Л. К.*, *Гариепи Р. В.* Теория меры и тонкие свойства функций. Новосибирск: Научная книга, 2002.

106. *Яковлев П. А. и др*. Алгоритмы локальной минимизации силового поля для трехмерных макромолекул // ЖВМ и МФ. 2019. Т. 59, № 12. С. 1994–2008.
107. *Abernethy J.*, *Hazan E.* Faster convex optimization: simulated annealing with an efficient universal barrier // arXiv:1507.02528.
108. *Agafonov A. D.* Lower bounds for conditional gradient type methods for minimizing smooth strongly convex functions // arXiv:2003.07073.
109. *Agarwal A.*, *Bottou L.* A lower bound for the optimization of finite sums // arXiv:1410.0723.
110. *Agarwal A.*, *Dekel O.*, *Xiao L.* Optimal algorithms for online convex optimization with multi-point bandit feedback // COLT. 2010. P. 28–40.
111. *Agarwal A. et al.* Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization // IEEE Transaction of Information. 2012. V. 58, № 5. P. 3235–3249.
112. *Agarwal N. et al*. Finding approximate local minima faster than gradient descent // Proceedings of the Forty-Ninth Annual ACM Symposium on the Theory of Computing, 2017.
113. *Agarwal N.*, *Hazan E.* Lower bounds for higher-order convex optimization // arXiv:1710.10329.
114. *Ahn K.* From proximal point method to Nesterov's acceleration // arXiv:2005.08304.
115. *Ahn K.*, *Sra S.* On tight convergence rates of without-replacement SGD // arXiv:2004.08657.
116. *Ajalloeian A.*, *Stich S.* Analysis of SGD with biased gradient estimators // Workshop on «Beyond First Order Methods in ML Systems» at the 37th International Conference on Machine Learning, Vienna, Austria, 2020.
117. *Akhavan A.*, *Pontil M.*, *Tsybakov A. B.* Exploiting Higher Order Smoothness in Derivative-free Optimization and Continuous Bandits // arXiv:2006.07862.
118. *Allen-Zhu Z. et al*. Even faster accelerated coordinate descent using non-uniform sampling // arXiv:1512.09103.
119. *Allen-Zhu Z. et al.* Much faster algorithm for matrix scaling // arXiv:1704.02315.
120. *Allen-Zhu Z.*, *Hazan E.* Optimal black-box reductions between optimization objectives // arXiv:1603.05642.
121. *Allen-Zhu Z.*, *Hazan E.* Variance reduction for faster non-convex optimization // arXiv:1603.05643.
122. *Allen-Zhu Z.* How to make the gradients small stochastically: even faster convex and nonconvex SGD // Advances in Neural Information Processing Systems. 2018. P. 1165–1175.
123. *Allen-Zhu Z.* Katyusha X: Practical momentum method for stochastic sum-of-nonconvex optimization // arXiv:1802.03866.
124. *Allen-Zhu Z.* Katyusha: the first direct acceleration of stochastic gradient methods // arXiv:1603.05953.

125. *Allen-Zhu Z.*, *Li Y.* Neon 2: Finding local minima via first-order oracle // arXiv:1711.06673.
126. *Allen-Zhu Z.* Natasha 2: Faster non-convex optimization than SGD // arXiv:1708.08694.
127. *Allen-Zhu Z.*, *Orecchia L.* Linear coupling: An ultimate unification of gradient and mirror descent // arXiv:1407.1537v4.
128. *Allen-Zhu Z.* Personal web page. URL: http://people.csail.mit.edu/zeyuan/
129. *Allen-Zhu Z.*, *Simchi-Levi D.*, *Wang X.* The Lingering of gradients: how to reuse gradients over time // Advances in Neural Information Processing Systems. 2018. P. 1252–1261.
130. *Allen-Zhu Z.*, *Yuan Y.*, *Sridharan K.* Exploiting the structure: stochastic gradient methods using raw clusters // arXiv:1602.02151.
131. *Altschuler J.*, *Weed J.*, *Rigollet P.* Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration // In Advances in Neural Information Processing Systems. 2017. P. 1964–1974.
132. *Anandkumar A.*, *Ge R.* Efficient approach for escaping higher order saddle points in non-convex optimization // arXiv:1602.05908.
133. *Andrei N.* 40 conjugate gradient algorithms for unconstrained optimization. A survey on their definition // ICI Technical Report № 13/08, March 14, 2008. URL: https://camo.ici.ro/neculai/p13a08.pdf
134. *Andrychowicz M. et al.* Learning to learn by gradient descent by gradient descent // arXiv:1606.04474.
135. *Anikin A. et al.* Modern efficient numerical approaches to regularized regression problems in application to traffic demands matrix calculation from link loads // Proceedings of International conference ITAS — 2015. Russia, Sochi, September, 2015. arXiv:1508.00858.
136. *Antipin A. S.* Gradient approach of computing fixed points of equilibrium problems // Journal of Global Optimization. 2002. V. 24, № 3. P. 285–309.
137. *Antonakopoulos K.*, *Belmega V.*, *Mertikopoulos P.* An adaptive Mirror-Prox method for variational inequalities with singular operators // Advances in Neural Information Processing Systems. 2019. P. 8455–8465.
138. *Arjevani Y. et al.* Lower bounds for non-convex stochastic optimization // arXiv:1912.02365.
139. *Arjevani Y. et al.* On the complexity of minimizing convex finite sums without using the indices of the individual functions // arXiv:2002.03273.
140. *Arjevani Y. et al.* Second-order information in non-convex stochastic optimization: power and limitations // arXiv:2006.13476.
141. *Arjevani Y.* Limitation on variance-reduction and acceleration schemes for finite sum optimization // arXiv:1706.01686.
142. *Arjevani Y.*, *Shamir O.* Communication complexity of distributed convex learning and optimization // Advances in neural information processing systems. 2015. P. 1756–1764.

143. *Arjevani Y.*, *Shamir O.*, *Shiff R.* On lower and upper bounds in smooth and strongly convex optimization // Journal of Machine Learning Research. 2016. V. 17. P. 1–51.
144. *Arjevani Y.*, *Shamir O.*, *Shiff R.* Oracle complexity of second-order methods for smooth convex optimization // arXiv:1705.07260.
145. *Asi H.*, *Duchi J. C.* The importance of better models in stochastic optimization // arXiv:1903.08619.
146. *Assran M.*, *Rabbat M.* On the convergence of Nesterov's accelerated gradient method in stochastic settings // arXiv:2002.12414.
147. *Aybat N. S. et al.* Robust accelerated gradient methods for smooth strongly convex functions // arXiv:1805.10579.
148. *Azizian W. et al.* A tight and unified analysis of extragradient for a whole spectrum of differentiable games // arXiv:1906.05945.
149. *Bach F.*, *Levy K. Y.* A universal algorithm for variational inequalities adaptive to smoothness and noise // arXiv:1902.01637.
150. *Baes M.* Estimate sequence methods: extensions and approximations // URL: http://www.optimization-online.org/DB_FILE/2009/08/2372.pdf
151. *Bansal N.*, *Gupta A.* Potential-function proofs for first-order methods // arXiv:1712.04581.
152. *Barre M.*, *Taylor A.*, *Bach F.* Principled analyses and design of first-order methods with inexact proximal operators // arxiv:2006.06041.
153. *Barre M.*, *Taylor A.*, *d'Aspremont A.* Complexity guarantees for Polyak steps with momentum // arXiv:2002.00915.
154. *Bartlett P. L.*, *Mendelson S.* Empirical minimization // Probability theory and related fields. 2006. V. 135, № 3. P. 311–334.
155. *Bauschke H. H.*, *Bolte J.*, *Teboulle M.* A descent Lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications // Mathematics of Operations Research. 2017. V. 42, № 2. P. 330–348.
156. *Bauschke H. H.*, *Borwein J. M.*, *Combettes P. L.* Bregman monotone optimization algorithms // SIAM J. of Control and Optimization. 2003. V. 42, № 2. P. 596–636.
157. *Bauschke H. H.*, *Combettes P. L.* Convex analysis and monotone operator theory in Hilbert spaces. Springer, 2017.
158. *Bayandina A. et al.* Mirror descent and convex optimization problems with non-smooth inequality constraints. In Lecture Notes in Mathematics 2227. Large scale and distributed optimization. Pontus Giselsson and Anders Rantzer (Eds.). 2018. P. 181–214.
159. *Baydin A. G. et al.* Automatic differentiation in machine learning: a survey // arXiv:1502.05767.
160. *Beck A.* First-order methods in optimization. MOS-SIAM Series on Optimization. SIAM, 2017.
161. *Beck A.* On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and

decomposition schemes // SIAM Journal on Optimization. 2015. V. 25, № 1. P. 185–209.

162. *Beck A.*, *Teboulle M.* A fast iterative shrinkage-thresholding algorithm for linear inverse problems // SIAM Journal on Imaging Sciences. 2009. V. 2. P. 183–202.
163. *Beck A.*, *Teboulle M.* Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems // IEEE Transactions on Image Processing. 2009. V. 18, № 11. P. 2419–2434.
164. *Ben-Tal A.*, *Nemirovski A.* Lectures on modern convex optimization analysis, algorithms, and engineering applications. Philadelphia: SIAM, 2019. URL: http://www2.isye.gatech.edu/~nemirovs/Lect_ModConvOpt.pdf
165. *Ben-Tal A.*, *Nemirovski A.* On polyhedral approximation of the second-order cone // Math. of Oper. Research. 2001. V. 26, № 2. P. 193–205.
166. *Bern M. et al.* Support-graph preconditioners // SIAM J. Matrix Anal. Appl. 2006. V. 27. P. 930–951.
167. *Bertsekas D. P.* Constrained optimization and Lagrange multipliers methods. Athena Scientific, 1996.
168. *Bertsekas D. P.* Convex optimization theory. Athena Scientific, 2009.
169. *Bertsekas D. P.*, *Nedic A.*, *Ozdaglar A. E.* Convex analysis and optimization. Belmont, Massachusetts: Athena Scientific, 2003.
170. *Bertsekas D. P.*, *Tsitsiklis J. N.* Parallel and distributed computation: numerical methods. Prentice-Hall International, 1989.
171. *Beznosikov A.*, *Gorbunov E.*, *Gasnikov A.* Derivative-free method for decentralized distributed non-smooth optimization // IFAC. 2020. arXiv:1911.10645.
172. *Birgin E. G. et al.* Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models // Math. Programming. 2017. Ser. A. V. 163, № 1–2. P. 359–368.
173. *Blanchet J. et al.* Towards optimal running times for optimal transport // arXiv:1810.07717.
174. *Blum L. et al.* Complexity and real computation. Springer, 2012.
175. *Bolte J. et al.* First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems // SIAM Journal on Optimization. 2018. V. 28, № 3. P. 2131–2151.
176. *Borgwardt K. H.* The Simplex Method: a probabilistic analysis. Springer Science & Business Media, 2012. V. 1.
177. *Bottou L.*, *Curtis F. E.*, *Nocedal J.* Optimization methods for large-scale machine learning // arXiv:1606.04838.
178. *Boucheron S.*, *Lugosi G.*, *Massart P.* Concentration inequalities: A nonasymptotic theory of independence. Oxford University Press, 2013.
179. *Boyd S. et al.* Distributed optimization and statistical learning via the alternating direction method of multipliers // Foundations and Trends in Machine Learning. 2010. V. 3, № 1. P. 1–122.

180. *Boyd S.*, *Parikh N.* Proximal algorithms // Foundations and Trends in Optimization. 2014. V. 1, № 3. P. 123–231.
181. *Boyd S.* Personal web-page. URL: http://stanford.edu/~boyd/
182. *Boyd S.*, *Vandenberghe L.* Convex optimization. Cambridge University Press, 2004. URL: https://web.stanford.edu/~boyd/cvxbook/
183. *Boyd S.*, *Vandenberghe L.* Introduction to applied linear algebra — vectors, matrices and least squares. Cambridge Univ. Press, 2018. URL: https://web.stanford.edu/~boyd/vmls/
184. *Braverman M. et al.* The gradient complexity of linear regression // arXiv:1911.02212.
185. *Brent R. P.* Algorithms for minimization without derivatives. Prentice-Hall, 1973.
186. *Bubeck S.* Convex optimization: algorithms and complexity // Foundations and Trends in Machine Learning. 2015. V. 8, № 3–4. P. 231–357.
187. *Bubeck S.*, *Eldan R.* The entropic barrier: a simple and optimal self-concordant barrier // arXiv:1412.1587.
188. *Bubeck S. et al.* Complexity of highly parallel non-smooth convex optimization // arXiv:1906.10655.
189. *Bubeck S. et al.* Near-optimal method for highly smooth convex optimization // arXiv:1812.08026.
190. *Bubeck S.* Introduction to online optimization. Lecture Notes. 2011. P. 1–86. URL: http://sbubeck.com/BubeckLectureNotes.pdf
191. *Bubeck S.*, *Mikulincer D.* How to trap a gradient flow // arXiv:2001.02968.
192. *Bullins B.*, *Lai K. A.* Higher-order methods for convex-concave min-max optimization and monotone variational inequalities // arXiv:2007.04528.
193. *Bullins B.*, *Peng R.* Higher-order accelerated methods for faster non-smooth optimization // arXiv:1906.01621.
194. *Carderera A.*, *Diakonikolas J.*, *Pokutta S.* Locally accelerated conditional gradients // arXiv:1906.07867.
195. *Carmon Y. et al.* “Convex until proven guilty”: Dimension-free acceleration of gradient descent to non-convex functions // arXiv:1705.02766.
196. *Carmon Y. et al.* Accelerated methods for non-convex optimization // arXiv: 1611.00756.
197. *Carmon Y. et al.* Lower bounds for finding stationary points I // arXiv: 1710.11606.
198. *Carmon Y. et al.* Lower bounds for finding stationary points II: First-order methods // arXiv:1711.00841.
199. *Carmon Y. et al.* Variance reduction for matrix games // arXiv:1907.02056.
200. *Cartis C.*, *Gould N. I.*, *Toint P. L.* On the complexity of steepest descent, Newton’s and regularized Newton’s methods for nonconvex unconstrained optimization problems // SIAM journal on optimization. 2010. V. 20, № 6. P. 2833–2852.

201. *Cauchy A.* Méthode générale pour la résolution des systemes d'équations simultanées // Comp. Rend. Sci. Paris. 1847. V. 25, № 1847. P. 536–538.
202. *Cevher V.*, *Becker S.*, *Schmidt M.* Convex optimization for Big Data: Scalable, randomized and parallel algorithms for big data analytics // IEEE Signal Processing Magazine. 2014. V. 31, № 5. P. 32–43.
203. *Chambolle A.*, *Pock T.* A first-order primal-dual algorithm for convex problems with applications to imaging // Journal of Math. Imaging & Vision. 2011. V. 40, № 1. P. 120–145.
204. *Chambolle A.*, *Tan P.*, *Vaiter S.* Accelerated alternating descent methods for Dykstra-like problems // Journal of Mathematical Imaging and Vision. 2017. V. 59, № 3. P. 481–497.
205. *Chang T.-H.*, *Hong M.*, *Wang X.* Multi-agent distributed optimization via inexact consensus ADMM // IEEE Transactions on Signal Processing. 2015. V. 63, № 2. P. 482–487.
206. *Chen G.*, *Teboulle M.* Convergence analysis of a proximal-like minimization algorithm using Bregman functions // SIAM J. Optim. 1993. V. 3. P. 538–543.
207. *Chen Y.*, *Lan G.*, *Ouyang Y.* Accelerated schemes for class of variational inequalities // Math. Programming. Ser. B. 2017. V. 165, № 1. P. 113–149.
208. *Chen Y.*, *Lan G.*, *Ouyang Y.* Optimal primal-dual methods for class of saddle point problems // SIAM J. Optim. 2014. V. 24, № 4. P. 1779–1814.
209. *Chen Y.*, *Orvieto A.*, *Lucchi A.* An accelerated DFO algorithm for finite-sum convex functions // arXiv:2007.03311.
210. *Chen Z.*, *Zhou Y.* Momentum with variance reduction for nonconvex composition optimization // arXiv:2005.07755.
211. *Cipra B. A.* The best of the 20th century: Editors name top 10 algorithms // SIAM news. 2000. V. 33, № 4. P. 1–2.
URL: `https://archive.siam.org/pdf/news/637.pdf`
212. *Cohen M. B.*, *Diakonikolas J.*, *Orecchia L.* On acceleration with noise-corrupted gradients // arXiv:1805.12591.
213. *Cohen M. B. et al.* Almost-linear-time algorithm for Markov chain and new spectral primitives for directed graphs // arXiv:1611.00755.
214. *Cohen M. B. et al.* Geometric median in nearly linear time // arXiv:1606.05225.
215. *Cohen M. B. et al.* Matrix scaling and balancing via box constrained Newton's method and interior point methods // In Chris Umans, editor, 58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15–17, 2017. P. 902–913. IEEE Computer Society, 2017.
216. *Cohen M. B.*, *Sidford A.*, *Tian K.* Relative Lipschitzness in Extragradient Methods and a Direct Recipe for Acceleration // arXiv:2011.06572.
217. *Conn A. B.*, *Gould N. I. M.*, *Toint P. L.* Trust region methods. Philadelphia: SIAM, 2000.
218. *Conn A. R.*, *Scheinberg K.*, *Vicente L. N.* Introduction to derivative-free optimization. Siam, 2009. V. 8.

219. *Cox B.*, *Juditsky A.*, *Nemirovski A.* Decomposition techniques for bilinear saddle point problems and variational inequalities with affine monotone operators on domains given by linear minimization oracles // arXiv:1506.02444v3.
220. *Cuturi M.*, *Peyre G.* A smoothed dual formulation for variational Wasserstein problems // SIAM J. Imaging Science. 2016. V. 9. P. 320–343.
221. *d'Aspremont A.* Smooth minimization with approximate gradient // SIAM Journal on Optimization. 2008. V. 19, № 3. P. 1171–1183.
222. *Dadush D.*, *Huiberts S.* A friendly smoothed analysis of the simplex method // Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing. ACM, 2018. P. 390–403.
223. *Dang C. D.*, *Lan G.* On the convergence properties of non-euclidian extragradient methods for variational inequalities with generalized monotone operators // Comput. Optim. Appl. 2015. V. 60. P. 227–310.
224. *Danilova M. et al.* Recent Theoretical Advances in Non-Convex Optimization // arXiv:2012.06188.
225. *Davis D.*, *Drusvyatskiy D.* Robust stochastic optimization with the proximal point method // arXiv:1907.13307.
226. *Davis D.*, *Yin W.* Convergence rate analysis of several splitting schemes // Splitting methods in communication, imaging, science, and engineering. Cham: Springer, 2016. P. 115–163.
227. *de Klerk E.*, *Glineur F.*, *Taylor A. B.* On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions // Optimization Letters. 2017. V. 11, № 7. P. 1185–1199.
228. *Defazio A.* A simple practical accelerated method for finite sums // Advances in neural information processing systems. 2016. P. 676–684.
229. *Defazio A.*, *Bottou L.* On the ineffectiveness of variance reduced optimization for deep learning // Advances in Neural Information Processing Systems. 2019. P. 1755–1765.
230. *Défossez A. et al.* On the Convergence of Adam and Adagrad // arXiv: 2003.02395.
231. *Deng Q.*, *Cheng Y.*, *Lan G.* Optimal adaptive and accelerated stochastic gradient descent // arXiv:1810.00553.
232. *Dennis J. E.*, *More J. J.* Quasi-Newton methods, motivation and theory // SIAM Review. 1977. V. 19, № 1. P. 46–89.
233. *Devolder O.* Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization: PhD thesis. CORE UCL, March 2013.
234. *Devolder O.*, *Glineur F.*, *Nesterov Y.* Double smoothing technique for large-scale linearly constrained convex optimization // SIAM J. Optim. 2012. V. 22, № 2. P. 702–727.
235. *Devolder O.*, *Glineur F.*, *Nesterov Yu.* First order methods of smooth convex optimization with inexact oracle // Math. Programming. Ser. A. 2014. V. 146, № 1–2. P. 37–75.

236. *Devolder O.*, *Glineur F.*, *Nesterov Yu.* First order methods with inexact oracle: the smooth strongly convex case // CORE Discussion Paper 2013/16. 2013. URL: https://www.uclouvain.be/cps/ucl/doc/core/documents/coredp2013_16web.pdf

237. *Diakonikolas J.*, *Guzmán C.* Lower bounds for parallel and randomized convex optimization // arXiv:1811.01903.

238. *Diakonikolas J.* Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities // arXiv:2002.08872.

239. *Diakonikolas J.*, *Orecchia L.* Alternating randomized block coordinate descent gradients // arXiv:1805.09185.

240. *Diakonikolas J.*, *Orecchia L.* The approximate gap technique: A unified approach to optimal first-order methods // arXiv:1712.02485.

241. *Doikov N.*, *Nesterov Y.* Contracting proximal methods for smooth convex optimization // arXiv:1912.07972.

242. *Doikov N.*, *Nesterov Y.* Convex optimization based on global lower second-order models // arXiv:2006.08518.

243. *Doikov N.*, *Nesterov Y.* Inexact tensor methods with dynamic accuracies // arXiv:2002.09403.

244. *Doikov N.*, *Nesterov Y.* Local convergence of tensor // CORE Discussion Papers. 2019/21. 2019. URL: https://dial.uclouvain.be/pr/boreal/object/boreal%3A223026/datastream/PDF_01/view

245. *Dongarra J.*, *Sullivan F.* Guest editors' introduction: The top 10 algorithms // Computing in Science & Engineering. 2000. V. 2, № 1. P. 22.

246. *Dragomir R. A. et al.* Optimal complexity and certification of Bregman first-order methods // arXiv:1911.08510.

247. *Drori Y.*, *Taylor A. B.* Efficient first-order methods for convex minimization: a constructive approach // arXiv:1803.05676.

248. *Drori Y.*, *Teboulle M.* An optimal variants of Kelley's cutting-plane method // Math. Programming. Ser. A. 2016. V. 160, № 1–2. P. 321–351.

249. *Drori Y.*, *Teboulle M.* Performance of first-order methods for smooth convex minimization: a novel approach // Math. Programming. Ser. A. 2014. V. 145, № 1–2. P. 451–482.

250. *Drori Y.* The exact information-based complexity of smooth convex minimization // J. Complexity. 2017. V. 39. P. 1–16.

251. *Drusvyatskiy D.* Course notes: Convex analysis and optimization. 2019. URL: https://sites.math.washington.edu/~ddrusv/crs/Math_516/bookwithindex.pdf

252. *Drusvyatskiy D.*, *Fazel M.*, *Roy S.* An optimal first order method based on optimal quadratic averaging // SIAM Journal on Optimization. 2018. V. 28, № 1. P. 251–271.

253. *Drusvyatskiy D.*, *Ioffe A. D.*, *Lewis A. S.* Nonsmooth optimization using taylor-like models: error bounds, convergence, and termination criteria // arXiv:1610.03446.
254. *Drusvyatskiy D.* The proximal point method revisited // arXiv:1712.06038.
255. *Du S. S.*, *Hu W.* Linear Convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity // arXiv:1802.01504.
256. *Duchi J.*, *Ruan F.* Stochastic methods for composite optimization problems // arXiv:1703.08570.
257. *Duchi J. C. et al.* Optimal rates for zero-order convex optimization: The power of two function evaluations // IEEE Transactions on Information Theory. 2015. V. 61, № 5. P. 2788–2806.
258. *Duchi J. C.* Introductory lectures on stochastic optimization // IAS/Park City Mathematics Series. 2016. URL: `http://stanford.edu/~jduchi/PCMI Convex/Duchi16.pdf`
259. *Dvinskikh D. et al.* Adaptive gradient descent for convex and non-convex stochastic optimization // IFAC. 2020. arXiv:1911.08380.
260. *Dvinskikh D. et al.* On dual approach for distributed stochastic convex optimization over networks // 2019 IEEE Conference on Decision and Control (CDC). IEEE, 2019. arXiv:1903.09844.
261. *Dvinskikh D. et al.* Parallel and Distributed algorithms for ML problems // arXiv:2010.09585.
262. *Dvinskikh D.*, *Gasnikov A. V.* Decentralized and parallelized primal and dual accelerated methods for stochastic convex programming problems // Journal of Ill-posed Inverse problems. 2021 (in print).
263. *Dvinskikh D.* Stochastic Approximation versus Sample Average Approximation for population Wasserstein barycenter calculation // arXiv:2001.07697.
264. *Dvurechensky P. et al.* Decentralize and Randomize: Faster Algorithm for Wasserstein Barycenters // NIPS. 2018. arXiv:1806.03915.
265. *Dvurechensky P. et al.* Generalized Mirror Prox: Solving variational inequalities with monotone operator, inexact oracle, and unknown Hölder parameters // JOTA. 2020 (accepted). arXiv:1806.05140.
266. *Dvurechensky P. et al.* Near-optimal tensor methods for minimizing the gradient norm of convex function // arXiv:1912.03381.
267. *Dvurechensky P.*, *Gasnikov A.*, *Gorbunov E.* An accelerated directional derivative method for smooth stochastic convex optimization // EJOR. 2020 (accepted). arXiv:1804.02394.
268. *Dvurechensky P.*, *Gasnikov A.*, *Gorbunov E.* An accelerated methods for derivative-free smooth stochastic convex optimization // SIOPT. 2020 (accepted). arXiv:1802.09022.
269. *Dvurechensky P.*, *Gasnikov A.*, *Kamzolov D.* Universal intermediate gradient method for convex problems with inexact oracle // Optimization Methods and Software. 2020. arXiv:1712.06036.

270. *Dvurechensky P.*, *Gasnikov A.*, *Kroshnin A.* Computational optimal transport: complexity by accelerated gradient descent is better than by Sinkhorn's algorithms // ICML. 2018. arXiv:1802.04367.
271. *Dvurechensky P.*, *Gasnikov A.* Stochastic intermediate gradient method for convex Problems with inexact stochastic oracle // JOTA. 2016. V. 171, № 1. P. 121–145.
272. *Dvurechensky P.*, *Gasnikov A.*, *Tiurin A.* Randomized similar triangles method: A unifying framework for accelerated randomized optimization methods (coordinate descent, directional search, derivative-free method) // arXiv:1707.08486.
273. *Dvurechensky P.* Gradient method with inexact oracle for composite non-convex optimization // arXiv:1703.09180.
274. *Dvurechensky P.*, *Nesterov Yu.* Global performance guarantees of second-order methods for unconstrained convex minimization // CORE Discussion paper. 2018/32.
275. *Dvurechensky P.*, *Nesterov Yu.*, *Spokoiny V.* Primal-dual methods for solving infinite-dimensional games // JOTA. 2015. V. 7, № 1. P. 23–51.
276. *Ene A.*, *Nguyen H. L.*, *Vladu A.* Adaptive Gradient Methods for Constrained Convex Optimization // arXiv:2007.08840.
277. *Fang C. et al.* Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator // Advances in Neural Information Processing Systems. 2018. P. 687–697.
278. *Fang H.*, *Fan Z.*, *Friedlander M. P.* Fast convergence of stochastic subgradient method under interpolation // ICLR, 2021.
279. *Fatkhullin I.*, *Polyak B.* Optimizing static linear feedback: Gradient method // arXiv:2004.09875.
280. *Fercoq O.*, *Qu Z.* Restarting accelerated gradient methods with a rough strong convexity estimate // arXiv:1609.07358.
281. *Fercoq O.*, *Richtarik P.* Universal coordinate descent and line search for parallel coordinate descent // e-print, 2014. URL: `https://perso.telecom-paristech.fr/ofercoq/UniversalCoordinateDescentBirminghamSep2014.pdf`
282. *Floudas C. A.*, *Pardalos P. M.* Encyclopedia of optimization. Kluwer Academic Publishers, 2009.
283. *Foster D. et al.* The complexity of making the gradient small in stochastic convex optimization // arXiv:1902.04686.
284. *Franca G.*, *Robinson D. P.*, *Vidal R.* ADMM and accelerated ADMM as continuous dynamical systems // arXiv:1805.06579.
285. *Franklin J.*, *Lorenz J.* On the scaling of multidimensional matrices // Linear Algebra and its applications. 1989. V. 114. P. 717–735.
286. *Friedman E. J.*, *Oren S. S.* The Complexity of resource allocation and price mechanisms under bounded rationality // Economic Theory. 1995. V. 6, № 2. P. 225–250.

287. *Gabow H. N.*, *Tarjan R. E.* Faster scaling algorithms for general graph matching problems // Journal of the ACM (JACM). 1991. V. 38, № 4. P. 815–853.
288. *Gall F. L.* Powers of tensor and fast matrix multiplication // Proceedings of the 39th international symposium on symbolic and algebraic computation. 2014. P. 296–303.
289. *Gallager R. G.*, *Humblet P. A.*, *Spira P. M.* A distributed algorithm for spanning trees // ACM Transactions on Programming Languages and systems (TOPLAS). 1983. V. 5, № 1. P. 66–77.
290. *Garey M. R.*, *Johnson D. S.* Computers and intractability. New York: W. H. Freeman, 2002. V. 29.
291. *Gasnikov A. V. et al.* Near Optimal Methods for Minimizing Convex Functions with Lipschitz p-th Derivatives // Conference on Learning Theory. 2019. P. 1392–1393.
292. *Gasnikov A. V. et al.* Convex optimization in Hilbert space with application to inverse problems // Appl. Numer. Math. 2017 (submitted). arXiv: 1703.00267.
293. *Ge R.*, *Lee J. D.*, *Ma T.* Matrix completion has no spurious local minimum // Advances in Neural Information Processing Systems. 2016. P. 2973–2981.
294. *Ghadimi S.*, *Feyzmahdavian H. R.*, *Johansson M.* Global convergence of heavy-ball method for convex optimization // arXiv:1412.7457.
295. *Ghadimi S.*, *Lan G.* Stochastic first-and zeroth-order methods for non-convex stochastic programming // SIAM Journal on Optimization. 2013. V. 23, № 4. P. 2341–2368.
296. *Ghadimi S.*, *Lan G.*, *Zhang H.* Generalized uniformly optimal methods for nonlinear programming. URL: `https://pwp.gatech.edu/guanghui-lan/publications/`
297. *Ghadimi S.*, *Liu H.*, *Zhang T.* Second-order methods with cubic regularization under inexact information // arXiv:1710.05782.
298. *Gidel G.*, *Jebara T.*, *Lacoste-Julien S.* Frank — Wolfe algorithms for saddle point problems // arXiv:1610.07797.
299. *Gilabert P.*, *Montoro G.*, *Bertran E.* On the Wiener and Hammerstein models for power amplifier predistortion // 2005 Asia-Pacific Microwave Conference Proceedings. IEEE, 2005. V. 2.
300. *Gilpin A.*, *Pena J.*, *Sandholm T.* First-order algorithms with χ convergence for ε-equilibrium in two-person zero-sum game // Math. Programming. Ser. A. 2012. V. 133, № 1–2. P. 279–298.
301. *Gladin E. et al.* Solving strongly convex-concave composite saddle point problems with a small dimension of one of the variables // arXiv:2010.02280.
302. *Godichon-Baggioni A.*, *Saadane S.* On the rates of convergence of Parallelized Averaged Stochastic Gradient Algorithms // arXiv:1710.07926.
303. *Goldstein T. et al.* Fast alternating direction optimization methods // SIAM Journal on Imaging Sciences. 2014. V. 7, № 3. P. 1588–1623.

304. *Golub G. H.*, *Van Loan C. F.* Matrix computations. Baltimor, MD, USA: John Hopkins University Press, 2012.
305. *Goodfelow I. J.*, *Vintasls O.*, *Saxe A. M.* Qualitatively characterizing neural network optimization problems // ICLR. 2015. arXiv:1412.6544.
306. *Gorbunov E.*, *Danilova M.*, *Gasnikov A.* Stochastic optimization with heavy-tailed noise via Accelerated Gradient Clipping // NeurIPS. 2020. arXiv: 2005.10785.
307. *Gorbunov E.*, *Dvinskikh D.*, *Gasnikov A.* Optimal decentralized distributed algorithms for stochastic convex optimization // arXiv:1911.07363.
308. *Gorbunov E. et al.* Recent theoretical advances in decentralized distributed convex optimization // arXiv:2011.13259.
309. *Gorbunov E.*, *Hanzely F.*, *Richtarik P.* A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent // arXiv:1905.11261.
310. *Gower R. M. et al.* Accelerated stochastic matrix inversion: general theory and speeding up BFGS rules for faster second-order optimization // arXiv:1802.04079.
311. *Gower R. M. et al.* Variance-reduced methods for machine learning // Proceedings of the IEEE. 2020. V. 108, № 11. P. 1968–1983.
312. *Gower R. M.*, *Richtarik P.* Randomized quasi-Newton updates and linearly convergent matrix inversion algorithm // arXiv:1602.01768.
313. *Grant M.*, *Boyd S.*, *Ye Y.* Disciplined convex programming. Chapter in Global Optimization: From theory to implementation // Nonconvex Optimization and its Applications / L. Liberti and N. Maculan (eds.). Springer, 2006. P. 155–210.
URL: `http://stanford.edu/~boyd/papers/disc_cvx_prog.html`
314. *Grapiglia G. N.*, *Nesterov Y.* On Inexact solution of auxiliary problems in tensor methods for convex optimization // arXiv:1907.13023.
315. *Grapiglia G. N.*, *Nesterov Y.* Tensor methods for finding approximate stationary points of convex functions // arXiv:1907.07053.
316. *Grapiglia G. N.*, *Nesterov Y.* Tensor methods for minimizing functions with Hölder continuous higher-order derivatives // arXiv:1904.12559.
317. *Grapiglia G. N.*, *Nesterov Yu.* Accelerated regularized Newton method for minimizing composite convex functions // CORE Discussion Papers. 2018/10. 2018.
318. *Grapiglia G. N.*, *Nesterov Yu.* Regularized Newton methods for minimizing functions with Hölder continuous Hessian // SIAM J. Optim. 2017. V. 27, № 1. P. 478–506.
319. *Grimmer B.* Convergence Rates for deterministic and stochastic subgradient methods without Lipschitz continuity // arXiv:1712.04104.
320. *Guiges V.*, *Juditsky A.*, *Nemirovski A.* Non-asymptotic confidence bounds for the optimal value of a stochastic program // arXiv:1601.07592.
321. *Güler O.* New proximal point algorithms for convex minimization // SIAM Journal on Optimization. 1992. V. 2, № 4. P. 649–664.

322. *Guminov S. et al.* A universal modification of the linear coupling method // Optimization methods and software. 2019. V. 34, №. 3. P. 560–577.
323. *Guminov S. et al.* Accelerated alternating minimization // arXiv:1906.03622.
324. *Guminov S., Gasnikov A.* Accelerated methods for α-weakly-quasi-convex optimization problems // arXiv:1710.00797.
325. *Gunasekar S., Woodworth B., Srebro N.* Mirrorless mirror descent: a more natural discretization of Riemannian gradient flow // arXiv:2004.01025.
326. *Guzman C., Nemirovski A.* On lower complexity bounds for large-scale smooth convex optimization // Journal of Complexity. 2015. V. 31. P. 1–14.
327. *Hachez Y.* Convex optimization over non-negative polynomials: structured algorithms and applications. CORE UCL, PhD Thesis. 2003. URL: `https://perso.uclouvain.be/paul.vandooren/ThesisHachez.pdf`
328. *Hanzely F., Richtárik P.* Fastest rates for stochastic mirror descent methods // arXiv:1803.07374.
329. *Hanzely F., Richtárik P.* One method to rule them all: Variance reduction for data, parameters and many new methods // arXiv:1905.11266.
330. *Hanzely F., Richtarik P., Xiao L.* Accelerated Bregman proximal gradient method for relatively smooth functions // arXiv:1808.03045.
331. *Hardt M., Ma T.* Identity matters in Deep Learning // arXiv:1611.04231.
332. *Harrach B., Jahn J., Potthast R.* Beyond the Bakushinskii veto: Regularising linear inverse problems without knowing the noise distribution // arXiv:1811.06721.
333. *Harvey N. J. A. et al.* Tight analysis for non-smooth stochastic gradient descent // arXiv:1812.05217.
334. *Hazan E.* Introduction to online convex optimization // Foundations and Trends in Optimization. 2016. V. 2, № 3–4. P. 157–325.
335. *Hazan E., Kale S.* Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization // JMLR. 2014. V. 15. P. 2489–2512.
336. *Hazan E.* Lecture notes: Optimization for Machine Learning // arXiv: 1909.03550.
337. *Hazan E., Levy K. Y., Shalev-Shwartz S.* Beyond convexity: Stochastic quasi-convex optimization // Advances in Neural Information Processing Systems. 2015. P. 1594–1602.
338. *Heinz H. Bauschke, Patrick L. Combettes.* Convex analysis and monotone operator theory in Hilbert spaces. Springer Science & Business Media, 2011.
339. *Hendrikx H., Bach F., Massoulié L.* Accelerated decentralized optimization with local updates for smooth and strongly convex objectives // arXiv: 1810.02660.
340. *Hendrikx H., Bach F., Massoulié L.* An optimal algorithm for decentralized finite sum optimization // arXiv:2005.10675.
341. *Hendrikx H., Bach F., Massoulié L.* Asynchronous accelerated proximal stochastic gradient for strongly convex distributed finite sums // arXiv:1901.09865.

342. *Hendrikx H.*, *Bach F.*, *Massoulié L.* Dual-free stochastic decentralized optimization with variance reduction // arXiv:2006.14384.
343. *Hendrikx H. et al.* Statistically preconditioned accelerated gradient method for distributed optimization // arXiv:2002.10726.
344. *Hien L. T. K.*, *Zhao R.*, *Haskell W. B.* An inexact primal-dual smoothing framework for large-scale non-bilinear saddle point problems // e-print, 2019. arXiv:1711.03669.
345. *Hinder O.*, *Sidford A.*, *Sohoni N. S.* Near-optimal methods for minimizing star-convex functions and beyond // arXiv:1906.11985.
346. *Hofmann T. et al.* Variance reduced stochastic gradient descent with neighbors // NIPS. 2015. P. 2296–2304.
347. *Hu B.*, *Lessard L.* Dissipativity theory for Nesterov's accelerated method // arXiv:1706.04381.
348. *Hu C.*, *Pan W.*, *Kwok J. T.* Accelerated gradient methods for stochastic optimization and online learning // NIPS. 2009.
349. *Huang J. et al.* Strassen's algorithm reloaded // SC'16: Proceedings of the Inter-national Conference for High Performance Computing, Networking, Storage and Analysis. IEEE, 2016. P. 690–701.
350. *Hubbard J.*, *Schleicher D.*, *Sutherland S.* How to find all roots of complex polynomials by Newton's method // Inventiones mathematicae. 2001. V. 146, № 1. P. 1–33.
351. *Iusem A. N. et al.* Variance-based extragradient methods with line search for stochastic variational inequalities // SIAM Journal on Optimization. 2019. V. 29, № 1. C. 175–206.
352. *Ivanova A. et al.* Adaptive catalyst for smooth convex optimization // arXiv:1911.11271.
353. *Ivanova A. et al.* Adaptive Mirror Descent for the Network Utility Maximization Problem // IFAC. 2020. arXiv:1911.07354.
354. *Ivanova A. et al.* Composite optimization for the resource allocation problem // Optimization Methods and Software. 2020. arXiv:1810.00595.
355. *Ivanova A. et al.* Oracle Complexity Separation in Convex Optimization // arXiv:2002.02706.
356. *Jaggi M.* Revisiting Frank — Wolfe: Projection-free sparse convex optimization // Proceedings of the 30th International Conference on Machine Learning. Atlanta, Georgia, USA, 2013.
357. *Jain P.*, *Kar P.* Non-convex optimization for machine learning // arXiv: 1712.07897.
358. *Jambulapati A.*, *Sidford A.*, *Tian K.* A Direct χ iteration parallel algorithm for optimal transport // arXiv:1906.00618.
359. *Ji K. et al.* Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization // arXiv:2006.06359.
360. *Jiang B.*, *Lin T.*, *Zhang S.* A unified adaptive tensor approximation scheme to accelerate composite convex optimization // arXiv:1811.02427.

361. *Jiang B.*, *Wang H.*, *Zhang S.* An optimal high-order tensor method for convex optimization // arXiv:1812.06557.
362. *Jin C. et al.* A Short note on concentration inequalities for random vectors with subGaussian norm // arXiv:1902.03736.
363. *Jin C.*, *Netrapalli P.*, *Jordan M. I.* Accelerated gradient descent escapes saddle points faster than gradient descent // arXiv:1711.10456.
364. *Jin Y.*, *Sidford A.* Efficiently Solving MDPs with Stochastic Mirror Descent // arXiv:2008.12776.
365. *Juditsky A.*, *Kwon J.*, *Moulines E.* Unifying mirror descent and dual averaging // arXiv:1910.13742.
366. *Juditsky A.*, *Nemirovski A.* First order methods for nonsmooth convex large-scale optimization, I, II. Optimization for Machine Learning. MIT Press, 2012.
367. *Juditsky A.*, *Nemirovski A.*, *Tauvel C.* Solving variational inequalities with stochastic Mirror — Prox algorithm // Stochastic Systems. 2011. V. 1, № 1. P. 17–58.
368. *Juditsky A.*, *Nemirovski A. S.* Large deviations of vector-valued martingales in 2-smooth normed spaces // arXiv:0809.0813.
369. *Juditsky A.*, *Nesterov Yu.* Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization // Stoch. System. 2014. V. 4, № 1. P. 44–80.
370. *Kabanikhin S. I.* Inverse and ill-posed problems. De Gruyter, 2012.
371. *Kadkhodaie M. et al.* Accelerated alternating direction method of multipliers // Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2015. P. 497–506.
372. *Kairouz P. et al.* Advances and open problems in federated learning // arXiv:1912.04977.
373. *Kakde S. M.*, *Shalev-Shwartz S.*, *Tewari A.* On the duality of strong convexity and strong smoothness: learning applications and matrix regularization. URL: http://ttic.uchicago.edu/~shai/papers/KakadeShalevTewari09.pdf
374. *Kamzolov D.*, *Gasnikov A.*, *Dvurechensky P.* On the optimal combination of tensor optimization methods // LNCS. 2020. V. 12422. P. 1–18. (N. Olenev et al. (Eds.): OPTIMA 2020). arXiv:2002.01004.
375. *Kamzolov D.*, *Gasnikov A.* Near-optimal hyperfast second-order method for convex optimization and its sliding // arXiv:2002.09050.
376. *Karimi H.*, *Nutini J.*, *Schmidt M.* Linear convergence of gradient and proximal-gradient methods under the Polyak — Lojasiewicz condition // arXiv: 1608.04636.
377. *Karimi S.*, *Vavasis S. A.* A unified convergence bound for conjugate gradient and accelerated gradient // arXiv:1605.00320.
378. *Karimireddy S. P. et al.* Scaffold: Stochastic controlled averaging for on-device federated learning // International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

379. *Kavis A. et al.* UniXGrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization // Advances in Neural Information Processing Systems. 2019. P. 6257–6266.
380. *Kelner J. A. et al.* A simple, combinatorial algorithm for solving SDD systems in nearly-linear time // Proceeding STOC'13. P. 911–920.
381. *Kerdreux T., d'Aspremont A., Pokutta S.* Restarting Frank — Wolfe // arXiv: 1810.02429.
382. *Khaled A. et al.* Unified analysis of stochastic gradient methods for composite convex and smooth optimization // arXiv:2006.11573.
383. *Khaled A., Mishchenko K., Richtarik P.* Tighter theory for local SGD on identical and heterogeneous data // arXiv:1909.04746.
384. *Kim D.* Accelerated proximal point method for maximally monotone operators // arXiv:1905.05149.
385. *Kim D., Fessler J. A.* Adaptive restart of the optimized gradient method for convex optimization // arXiv:1703.04641.
386. *Kim D., Fessler J. A.* Generalizing the optimized gradient method for smooth convex minimization // arXiv:1607.06764.
387. *Kim D., Fessler J. A.* Optimized first-order methods for smooth convex optimization // Math. Programming. Ser. A. 2016. V. 159, № 1–2. P. 81–107.
388. *Kim D., Fessler J. A.* Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions // arXiv:1803.06600.
389. *Kinderlehrer D., Stampacchia G.* An introduction to variational inequalities and their applications. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM). (Classics in Applied Mathematics; V. 31).
390. *Koloskova A. et al.* A unified theory of decentralized SGD with changing topology and local updates // arXiv:2003.10422.
391. *Konecny J. et al.* Federated optimization: distributed machine learning for one-device intelligent // arXiv:1610.02527.
392. *Kornowski G., Shamir O.* High-Order Oracle Complexity of Smooth and Strongly Convex Optimization // arXiv:2010.06642.
393. *Kotsalis G., Lan G., Li T.* Simple and optimal methods for stochastic variational inequalities, I: operator extrapolation // arXiv:2011.02987.
394. *Krawtschenko R. et al.* Distributed Optimization with Quantization for Computing Wasserstein Barycenters // arXiv:2010.14325.
395. *Kroshnin A. et al.* On the complexity of approximating Wasserstein barycenter // ICML. 2019. arXiv:1901.08686.
396. *Kuhn H. W.* The Hungarian method for the assignment problem // Naval research logistics quarterly. 1955. V. 2, № 1–2. P. 83–97.
397. *Kulakova A., Danilova M., Polyak B.* Non-monotone behavior of the heavy ball method // arXiv:1811.00658.
398. *Kulunchakov A., Mairal J.* Estimate sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise // arXiv:1901.08788.

399. *Kulunchakov A., Mairal J. A.* Generic acceleration framework for stochastic composite optimization // arXiv:1906.01164.
400. *Lacost-Julien S., Schmidt M., Bach F.* A simpler approach to obtaining χ convergence rate for the projected stochastic subgradient method // arXiv: 1212.2002v2.
401. *Lacoste-Julien S., Jaggi M.* On the global linear convergence of Frank — Wolfe optimization variants // In Advances in Neural Information Processing Systems. 2015. P. 496–504.
402. *Lan G.* Bundle-level type methods uniformly optimal for smooth and non-smooth convex optimization // Math. Programming. Ser. A. 2015. V. 149, № 1–2. P. 1–45.
403. *Lan G.* First-order and stochastic optimization methods for Machine Learning. Springer, 2020. URL: `https://pwp.gatech.edu/guanghui-lan/`
404. *Lan G.* Gradient sliding for composite optimization // Math. Programming. Ser. A and B. 2016. V. 159, № 1–2. P. 201–235.
405. *Lan G., Lee S., Zhou Y.* Communication-efficient algorithms for decentralized and stochastic optimization // arXiv:1701.03961.
406. *Lan G., Li Z., Zhou Y.* A unified variance-reduced accelerated gradient method for convex optimization // arXiv:1905.12412.
407. *Lan G., Monteiro R. D. C.* Iteration-complexity of first-order augmented Lagrangian methods for convex programming // Math. Programming. Ser. A. 2016. V. 155, № 1–2. P. 511–547.
408. *Lan G., Monteiro R. D. C.* Iteration-complexity of first-order penalty methods for convex programming // Math. Programming. Ser. A. 2013. V. 138, № 1–2. P. 115–139.
409. *Lan G., Nemirovski A., Shapiro A.* Validation analysis of mirror descent stochastic approximation method // Mathematical programming. 2012. V. 134, № 2. P. 425–458.
410. *Lan G.* Personal web-page. URL: `https://pwp.gatech.edu/guanghui-lan/`
411. *Lan G., Romeijn E., Zhou Z.* Conditional Gradient Methods for convex optimization with function constraints // arXiv:2007.00153.
412. *Lan G., Yang Y.* Accelerated stochastic algorithms for non-convex finite-sum and multi-block optimization // arXiv:1805.05411.
413. *Lan G., Zhou Y.* An optimal randomized incremental gradient method // arXiv:1507.02000.
414. *Lan G., Zhou Y.* Randomized gradient extrapolation for distributed and stochastic optimization. URL: `https://pwp.gatech.edu/guanghui-lan/publications/`
415. *Larson J., Menickelly M., Wild S. M.* Derivative-free optimization methods // Acta Numerica. 2019. V. 28. P. 287–404.
416. *Le Roux N., Schmidt M., Bach F.* A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite

training sets // In Advances in Neural Information Processing Systems (NIPS). 2012. arXiv:1202.6258.

417. *Lee J. D. et al.* First-order methods almost always avoid saddle point // arXiv:1710.07406.
418. *Lee J. D. et al.* Gradient descent only converges to minimizers // 29th Annual Conference on Learning Theory (COLT). 2016. P. 1246–1257.
419. *Lee Y.-T., Sidford A., Wong S. C.-W.* A faster cutting plane method and its implications for combinatorial and convex optimization // arXiv:1508.04874.
420. *Lee Y. T., Sidford A.* Path-finding methods for linear programming: Solving linear programs in χ iterations and faster algorithms for maximum flow // In 55th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2014, 18–21 October, 2014. Philadelphia, PA, USA. P. 424–433.
421. *Lee Y. T., Sidford A., Vempala S. S.* Efficient convex optimization with membership oracles // Conference On Learning Theory. PMLR, 2018. P. 1292–1294.
422. *Lee Y. T., Sidford A., Vempala S. S.* Efficient convex optimization with oracles // Building Bridges II. Berlin, Heidelberg: Springer, 2019. P. 317–335.
423. *Lei Y., Tang K.* Stochastic composite mirror descent: optimal bounds with high probabilities // Advances in Neural Information Processing Systems. 2018. P. 1526–1536.
424. *Lemarechal C., Nemirovskii A., Nesterov Yu.* New variants of bundle methods // Math. prog. 1995. V. 69. P. 111–148.
425. *Lemarechal C., Sagaztizabal C.* Practice aspects of Moreau — Yosida regularization: Theoretical preliminaries // SIAM Journal on Optimization. 1997. V. 7, № 2. P. 367–385.
426. *Levy K. Y., Yurtsever A., Chevher V.* Online adaptive methods, universality and acceleration // arXiv:1809.02864.
427. *Li C. J. et al.* ROOT-SGD: Sharp Nonasymptotics and Asymptotic Efficiency in a Single Algorithm // arXiv:2008.12690.
428. *Li H. et al.* A sharp convergence rate analysis for distributed accelerated gradient methods // arXiv:1810.01053.
429. *Li H., Lin Z., Fang Y.* Optimal accelerated variance reduced extra and diging for strongly convex and smooth decentralized optimization // arXiv: 2009.04373.
430. *Li H., Lin Z.* Revisiting EXTRA for smooth distributed optimization // arXiv: 2002.10110.
431. *Li Z. et al.* PAGE: A Simple and Optimal Probabilistic Gradient Estimator for Nonconvex Optimization // arXiv:2008.10898.
432. *Liang J., Monteiro R. D. C.* A doubly accelerated inexact proximal point method for nonconvex composite optimization problems // arXiv:1811.11378.
433. *Lin H., Mairal J., Harchaoui Z.* A universal catalyst for first-order optimization // Proceedings of 29th International conference Neural Information Processing Systems (NIPS). Montreal, Canada. December, 7–12, 2015. P. 9.

URL: https://papers.nips.cc/paper/5928-a-universal-catalyst-for-first-order-optimization.pdf

434. *Lin H.*, *Mairal J.*, *Harchaoui Z.* Catalyst acceleration for first-order convex optimization: from theory to practice // arXiv:1712.05654.
435. *Lin T.*, *Jin C.*, *Jordan M.* Near-optimal algorithms for minimax optimization // arXiv:2002.02417.
436. *Lin T.*, *Jordan M.* A Control-theoretic perspective on optimal high-order optimization // arXiv:1912.07168.
437. *Lin Z.*, *Li H.*, *Fang C.* Accelerated optimization for Machine Learning. Springer, 2020.
438. *Loizou N.*, *Richtarik P.* Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods // arXiv: 1712.09677.
439. *Lu H.*, *Freund R. M.*, *Nesterov Yu.* Relatively-smooth convex optimization by first order methods, and applications // arXiv:1610.05708.
440. *Lucchi A.*, *Kohler J.* A stochastic tensor method for non-convex optimization // arXiv:1911.10367.
441. *Mai V. V.*, *Johansson M.* Anderson acceleration of proximal gradient methods // arXiv:1910.08590.
442. *Mairal J.* Optimization with first-order surrogate functions // ICML. 2013. arXiv:1305.3120.
443. *Malitsky Y.*, *Mishchenko K.* Adaptive gradient descent without descent // arXiv:1608.08883.
444. *Malitsky Y.*, *Pock T.* A first-order primal-dual algorithm with linesearch // arXiv:1608.08883.
445. *Malitsky Y.* Proximal extrapolated gradient methods for variational inequalities // Optimization Methods and Software. 2018. V. 33, № 1. P. 140–164.
446. *Mason J. C.*, *Handscomb D. C.* Chebyshev polynomials. Chapman and Hall/CRC, 2002.
447. *McMahan B. H. et al.* Communication-efficient learning of deep networks from decentralized optimization // arXiv:1602.05629.
448. *Mishchenko K. et al.* A delay-tolerant proximal-gradient algorithm for distributed learning // International Conference on Machine Learning. 2018. P. 3584–3592.
449. *Mishchenko K.*, *Khaled A.*, *Richtarik P.* Random reshuffling: Simple analysis with vast improvements // arXiv:2006.05988.
450. *Monteiro R.*, *Svaiter B.* An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods // SIAM Journal on Optimization. 2013. V. 23, № 2. P. 1092–1125.
451. *Mordukhovich B. S.* Variational analysis and applications. Springer, 2018.
452. *Mordukhovich B. S.* Variational analysis and generalized differentiation. I basic theory, II applications. Comprehensive studies in mathematics. Springer, 2006.

453. *Morin M.*, *Giselsson P.* Sampling and update frequencies in proximal variance reduced stochastic gradient methods // arXiv:2002.05545.

454. *Moulines E.*, *Bach F. R.* Non-asymptotic analysis of stochastic approximation algorithms for machine learning // Advances in Neural Information Processing Systems. 2011. P. 451–459.

455. *Muehlebach M.*, *Jordan M. I.* Optimization with momentum: Dynamical, control-theoretic, and symplectic perspectives // arXiv:2002.12493.

456. *Murty K. G.*, *Kabadi S. N.* Some NP-complete problems in quadratic and nonlinear programming // Mathematical programming. 1987. V. 39. P. 117–129.

457. *Narkiss G.*, *Zibulevsky M.* Sequential subspace optimization method for large-scale unconstrained problems // Tech. report CCIT № 559, EE Dept. Technion, 2005. URL: https://ie.technion.ac.il/~mcib/sesop_report_version301005.pdf

458. *Necoara I.*, *Nesterov Y.*, *Glineur F.* Linear convergence of first order methods for non-strongly convex optimization // Math. Programming. Ser. A. 2019. V. 175, № 1–2. P. 69–107.

459. *Nedic A.*, *Olshevsky A.*, *Shi W.* Achieving geometric convergence for distributed optimization over time-varying graphs // arXiv:1607.03218.

460. *Nedic A.*, *Ozdaglar A.* Cooperative distributed multi-agent optimization. In Convex Optimization in Signal Processing and Communications. Cambridge Univ. Press, 2009. P. 340–386.

461. *Nemirovski A.* Advanced Nonlinear Programming // Lectures, ISyE 7683 Spring 2019. URL: https://www2.isye.gatech.edu/~nemirovs/Trans_ModConvOpt.pdf

462. *Nemirovski A.* Information-based complexity of convex programming. Technion, Fall Semester 1994/95. URL: http://www2.isye.gatech.edu/~nemirovs/Lec_EMCO.pdf

463. *Nemirovski A.* Information-based complexity of linear operator equations // Journal of Complexity. 1992. V. 8. P. 153–175.

464. *Nemirovski A.* Introduction to Linear Optimization // Lectures, ISyE 6661 Fall 2018. URL: https://www2.isye.gatech.edu/~nemirovs/OPTI_Transparencies.pdf

465. *Nemirovski A.*, *Onn S.*, *Rothblum U. G.* Accuracy certificates for computational problems with convex structure // Math. of Operation Research. 2010. V. 35, № 1. P. 52–78.

466. *Nemirovski A.* Prox-method with rate of convergence χ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems // SIAM Journal on Optimization. 2004. V. 15. P. 229–251.

467. *Nemirovski A.* What can be expressed via conic quadratic and semidefinite programming // Presentation. Technion. 1998. URL: https://www2.isye.gatech.edu/~nemirovs/CONRUT.pdf

468. *Nesterov Yu.* Accelerating the cubic regularization of Newton's method on convex problems // Math. Programming. Ser. B. 2008. V. 112, № 1. P. 159–181.

469. *Nesterov Yu.* Efficiency of coordinate descent methods on large scale optimization problems // SIAM Journal on Optimization. 2012. V. 22, № 2. P. 341–362.

470. *Nesterov Yu. et al.* Primal-dual accelerated gradient descent with line search for convex and nonconvex optimization problems // Optimization methods and software. 2020. arXiv:1809.05895.

471. *Nesterov Yu.* Gradient methods for minimizing composite functions // Math. Programming. Ser. B. 2013. V. 140, № 1. P. 125–161.

472. *Nesterov Yu.* How to make the gradients small // Proc. of OPTIMA 88. 2012. P. 10–11.

473. *Nesterov Yu.* Implementable tensor methods in unconstrained convex optimization // CORE Discussion Papers. 2018/5. 22 p.

474. *Nesterov Yu.* Inexact accelerated high-order proximal-point methods // CORE Discussion Papers. 2020/8. 21 p.

475. *Nesterov Yu.* Inexact basic tensor methods // CORE Discussion Papers. 2019/23. 2019.

476. *Nesterov Yu.* Inexact high-order proximal-point methods with auxiliary search procedure // CORE Discussion paper. 2020/10. 23 p.

477. *Nesterov Yu.* Lectures on convex optimization. Springer, 2018.

478. *Nesterov Yu.* Lexicographical differentiation of nonsmooth functions // Math. Prog. 2005. V. 104, № 2–3. P. 669–700.

479. *Nesterov Yu.* Minimizing functions with bounded variation of subgradients // CORE Discussion Papers. 2005/79. 2005.

480. *Nesterov Yu.*, *Polyak B.* Cubic regularization of Newton method and its global performance // Math. Programming. Ser. A. 2006. V. 108, № 1. P. 177–205.

481. *Nesterov Yu.* Primal-dual subgradient methods for convex problems // Math. Programming. Ser. B. 2009. V. 120, № 1. P. 261–283.

482. *Nesterov Yu.*, *Shikhman V.* Distributed price adjustment based on convex analysis // Journal Optimization Theory and Applications. 2017. V. 172, № 2. P. 594–622.

483. *Nesterov Yu.*, *Shikhman V.* Dual subgradient method with averaging for optimal resource allocation // CORE Discussion paper. 2017/13. 19 p.

484. *Nesterov Yu.*, *Shpirko S.* Primal-dual subgradient method for huge-scale linear conic problem // SIAM Journal on Optimization. 2014. V. 24, № 3. P. 1444–1457.

485. *Nesterov Yu.* Smooth minimization of non-smooth function // Math. Programming. Ser. A. 2005. V. 103, № 1. P. 127–152.

486. *Nesterov Yu.* Soft clustering by convex electoral model // CORE Discussion paper. 2018/01.

487. *Nesterov Yu.*, *Spokoiny V.* Random gradient-free minimization of convex functions // Foundations of Computational Mathematics. 2017. V. 17, № 2. P. 527–566.
488. *Nesterov Yu.*, *Stich S.* Efficiency of accelerated coordinate descent method on structured optimization problems // SIAM J. Optim. 2017. V. 27, № 1. P. 110–123.
489. *Nesterov Yu.* Structure of non-negative polynomials and optimization problems // CORE Discussion papers № 1997049. 1997.
490. *Nesterov Yu.* Subgradient method for convex functions with nonstandard growth properties. Workshop "Three oracles", 2016. URL: http://www.mathnet.ru/php/seminars.phtml?option_lang=rus&presentid=16179
491. *Nesterov Yu.* Subgradient methods for huge-scale optimization problems // Math. Programming. Ser. A. 2013. V. 146, № 1–2. P. 275–297.
492. *Nesterov Yu.* Universal gradient methods for convex optimization problems // Math. Programming. Ser. A. 2015. V. 152, № 1–2. P. 381–404.
493. *Newton D.*, *Yousefian F.*, *Pasupathy R.* Stochastic gradient descent: Recent trends // Recent Advances in Optimization and Modeling of Contemporary Problems. INFORMS, 2018. P. 193–220.
494. *Niu F. et al.* HOGWILD! A lock-free approach to parallelizing stochastic gradient // arXiv:1106.5730.
495. *Nocedal J.*, *Wright S.* Numerical optimization. Springer, 2006.
496. *O'Donoghue B.*, *Candes E.* Adaptive restart for accelerated gradient schemes // Foundations of Computational Mathematics. 2015. V. 15. P. 715–732.
497. *Ochs P.*, *Fadili J.*, *Brox T.* Non-smooth non-convex Bregman minimization: unification and new algorithms // arXiv:1707.02278.
498. *Ogaltsov A.*, *Tyurin A.* Heuristic adaptive fast gradient method in stochastic optimization tasks // arXiv:1910.04825.
499. *Olshanskii M. A.*, *Tyrtyshnikov E. E.* Iterative methods for linear systems. SIAM, 2014.
500. *Orabona F.* A modern introduction to online learning // arXiv:1912.13213.
501. *Ouyang Y. et al.* An accelerated linearized direction method of multipliers // SIAM J. Imaging Science. 2015. V. 8, № 1. P. 644–681.
502. *Ouyang Y.*, *Xu Y.* Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems // arXiv:1808.02901.
503. *Palaniappan B.*, *Bach F.* Stochastic variance reduction methods for saddle-point problems // Advances in Neural Information Processing Systems. 2016. P. 1416–1424.
504. *Pasechnyk D. A.*, *Stonyakin F. S.* One method for minimization a convex Lipchitz-continuous function of two variables on a fixed square // arXiv: 1812.10300.
505. *Pedregosa F.*, *Scieur D.* Average-case acceleration through spectral density estimation // International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

506. *Pele O.*, *Werman M.* Fast and robust earth mover's distances // 2009 IEEE 12th International Conference on Computer Vision. IEEE, 2009. P. 460–467.
507. *Peypouquet J.* Convex optimization in normed spaces: theory, methods and examples. Springer, 2015. URL: http://web.dim.uchile.cl/~jpeypou/user/pages/03.publications/Pey_BOOK_2015
508. *Peyre G.*, *Cuturi M.* Computational optimal transport // arXiv:1803.00567.
509. *Pham N. H. et al.* ProxSARAH: an efficient algorithmic framework for stochastic composite nonconvex optimization // arXiv:1902.05679.
510. *Pletnev N.* Fast adaptive by constants of strong-convexity and Lipschitz for gradient first order methods // arXiv:2009.03971.
511. *Poljak B. T.* Iterative algorithms for singular minimization problems // Nonlinear Programming. V. 4. New York — London: Academic Press, 1981. P. 147–166.
512. *Polyak B.* History of mathematical programming in the USSR: analyzing the phenomenon // Math. Programming. Ser. B. 2002. V. 91, № 3. P. 401–416. URL: lab7.ipu.ru/files/polyak/Pol-rus-Baikal%2708.pdf
513. *Polyak B.*, *Tremba A.* Solving underdetermined nonlinear equations by Newton-like method // arXiv:1703.07810.
514. *Polyak B. T.*, *Juditsky A. B.* Acceleration of stochastic approximation by averaging // SIAM J. Control Optim. 1992. V. 30. P. 838–855.
515. *Polyak B. T.*, *Shcherbakov P. S.*, *Smirnov G.* Peack effects in stable linear difference equations // arXiv:1803.00808.
516. *Qian X. et al.* SGD with arbitrary sampling: general analysis and improved rates // International Conference on Machine Learning. 2019. P. 5200–5209.
517. *Quanrud K.* Approximating optimal transport with linear programs // arXiv: 1810.05957.
518. *Rajput S.*, *Gupta A.*, *Papailiopoulos D.* Closing the convergence gap of SGD without replacement // arXiv:2002.10400.
519. *Rakhlin A.*, *Shamir O.*, *Sridharan K.* Making gradient descent optimal for strongly convex stochastic optimization // Proceedings of the 29th International Conference on Machine Learning (ICML). Edinburgh, Scotland. June, 26 – July, 1, 2012. P. 8. URL: http://icml.cc/2012/papers/261.pdf
520. *Rakhlin A.*, *Sridharan K.* Statistical learning theory and sequential prediction. Lecture Notes in University of Pennsylvania. 2014. URL: http://www.mit.edu/~rakhlin/courses/stat928/stat928_notes.pdf
521. *Reddi S. J.*, *Kale S.*, *Kumar S.* On the convergence of Adam and beyond // arXiv:1904.09237.
522. *Renegar J.* A mathematical view of interior-point methods in convex optimization. MPS-SIAM Series on Optimization, 2001.
523. *Renegar J.*, *Grimmer B.* A simple nearly-optimal restart scheme for speeding-up first order methods // arXiv:1803.00151.
524. *Richtarik P.* Personal web page. URL: http://www.maths.ed.ac.uk/~prichtar/

525. *Richtarik P.*, *Takac M.* Stochastic reformulation of linear systems: algorithms and convergence theory // arXiv:1706.01108.
526. *Risteski A.*, *Li Y.* Algorithms and matching lower bounds for approximately-convex optimization // Advances in Neural Information Processing Systems. 2016. P. 4745–4753.
527. *Rockafellar R. T.* Convex analysis. Princeton University Press, 1996.
528. *Rodomanov A.*, *Nesterov Y.* Greedy quasi-Newton methods with explicit superlinear convergence // arXiv:2002.00657.
529. *Rodomanov A.*, *Nesterov Y.* New results on superlinear convergence of classical quasi-Newton methods // arXiv:2004.14866.
530. *Rogozin A. et al.* Towards accelerated rates for distributed optimization over time-varying networks // arXiv:2009.11069.
531. *Rogozin A.*, *Gasnikov A.* Projected gradient method for decentralized optimization over time-varying networks // LNCS. 2020 V. 12422. P. 1–18. (N. Olenev et al. (Eds.): OPTIMA 2020). arXiv:1911.08527.
532. *Roulet V.*, *d'Aspremont A.* Sharpness, restart and acceleration // NIPS. 2017. P. 1119–1129.
533. *Ruder A.* An overview of gradient descent optimization algorithms // arXiv: 1609.04747.
534. *Ruppert D.* Efficient estimations from a slowly convergent Robbins — Monro process. Cornell University Operations Research and Industrial Engineering, 1988.
535. *Ryabtsev A.* Error accumulation in conjugate gradient method for degenerate problem // arXiv:2004.10242.
536. *Saad Y.* Iterative methods for sparse linear systems. SIAM, 2003. URL: `https://www-users.cs.umn.edu/~saad/IterMethBook_2ndEd.pdf`
537. *Sadiev A.*, *Beznosikov A.*, *Gasnikov A.* Gradient-free methods with inexact oracle for convex-concave stochastic saddle-point problem // International Conference on Mathematical Optimization Theory and Operations Research. Cham: Springer, 2020. URL: `https://arxiv.org/pdf/2005.05913.pdf`
538. *Sadiev A. et al.* Zeroth-Order Algorithms for Smooth Saddle-Point Problems // arXiv:2009.09908.
539. *Scaman K. et al.* Optimal algorithms for non-smooth distributed optimization in networks // arXiv:1806.00291.
540. *Scaman K. et al.* Optimal algorithms for smooth and strongly convex distributed optimization in networks // arXiv:1702.08704.
541. *Scieur D.*, *d'Aspremont A.*, *Bach F.* Regularized nonlinear acceleration // arXiv:1606.04133.
542. *Scieur D. et al.* Integration methods and accelerated optimization algorithms // arXiv:1702.06751.
543. *Scieur D.*, *Pedregosa F.* Universal average-case optimality of Polyak momentum // arXiv:2002.04664.

544. *Scoy B. V.*, *Freeman R. A.*, *Lynch K. M.* The fastest known globally convergent first-order method for the minimization of strongly convex function // IEEE Control Systems Letter. 2018. V. 2, № 1. P. 49–54.
545. *Scutari G. et al.* Convex optimization, game theory, and variational inequality theory // IEEE Signal Processing Magazine. 2010. V. 27, № 3. P. 35–49.
546. *Shafi S. Y.*, *Arcak M.*, *Ghaoui L. E.* Graph weight allocation to meet Laplacian spectral constraints // IEEE Trans. on Automatic Control. 2012. V. 57, № 7. P. 1872–1877.
547. *Shalev-Shwartz S.*, *Ben-David S.* Understanding Machine Learning: From theory to algorithms. Cambridge University Press, 2014.
548. *Shalev-Shwartz S. et al.* Learnability, stability and uniform convergence // Journal of Machine Learning Research. 2010. V. 11. P. 2635–2670.
549. *Shalev-Shwartz S. et al.* Stochastic Convex Optimization // COLT. 2009. URL: `https://ttic.uchicago.edu/~karthik/nonlinearTR.pdf`
550. *Shalev-Shwartz S.*, *Zhang T.* Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization // ICML. 2014. P. 64–72.
551. *Shamir O.* An optimal algorithm for bandit and zero-order convex optimization with two-point feedback // Journal of Machine Learning Research. 2017. V. 18. P. 1–11.
552. *Shamir O.* On the complexity of bandit and derivative-free stochastic convex optimization // Conference on Learning Theory. 2013. P. 3–24.
553. *Shapiro A.*, *Dentcheva D.*, *Ruszczynski A.* Lecture on stochastic programming: Modeling and theory. MOS SIAM Series on Optimization, 2014.
554. *Shapiro A.*, *Nemirovski A.* On complexity of stochastic programming problems. Continuous: Current trends and applications. Eds. V. Jeyakumar and A. Rubinov. Springer, 2005. P. 111–144.
555. *Shi B.*, *Su W. J.*, *Jordan M. I.* On learning rates and Schrodinger operators // arXiv:2004.06977.
556. *Simchowitz M.*, *El Alaoui A.*, *Recht B.* Tight query complexity lower bounds for PCA via finite sample deformed Wigner law // Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing. ACM, 2018. P. 1249–1259.
557. *Simchowitz M.* On the randomized complexity of minimizing a convex quadratic function // arXiv:1807.09386.
558. *Smale S.* On the average number of steps of the simplex method of linear programming // Mathematical programming. 1983. V. 27, № 3. P. 241–262.
559. *Song C.*, *Ma Y.* Towards unified acceleration of high-order algorithms under Hölder continuity and uniform convexity // arXiv:1906.00582.
560. *Song L. et al.* On the complexity of learning neural networks // Advances in Neural Information Processing Systems. 2017. P. 5514–5522.
561. *Spall J. C.* Introduction to stochastic search and optimization: estimation, simulation and control. Wiley, 2003.

562. *Spielman D. A.*, *Teng S.-H.* Nearly-linear time algorithm for preconditioning and solving symmetric, diagonally dominated linear systems // SIAM J. Matrix Anal. & Appl. 2014. V. 35, № 3. P. 835–885.
563. *Spielman D. A.*, *Teng S. H.* Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time // Journal of the ACM (JACM). 2004. V. 51, № 3. P. 385–463.
564. *Sridharan K.* Learning from an optimization viewpoint. PhD Thesis, Toyota Technological Institute at Chicago, 2011. URL: http://ttic.uchicago.edu/~karthik/thesis.pdf
565. *Stich S.*, *Karimireddy P.* The error-feedback framework: Better rates for SGD with delayed gradients and compressed communications // arXiv:1909.05350.
566. *Stich S. U.* Unified optimal analysis of the (stochastic) gradient method // arXiv:1907.04232.
567. *Stonyakin F. et al.* Gradient methods for problems with inexact model of the objective // International Conference on Mathematical Optimization Theory and Operations Research. Cham: Springer, 2019. P. 97–114.
568. *Stonyakin F. et al.* Inexact model: a framework for optimization and variational inequalities // arXiv:1902.00990.
569. *Stonyakin F. et al.* Inexact relative smoothness and strong convexity for optimization and variational inequalities by inexact model // arXiv:2001.09013.
570. *Stonyakin F. et al.* Mirror descent for constrained optimization problems with large subgradient values values of functional constraints // Computer Research and Modeling. 2020. V. 12, № 2. P. 301–317.
571. *Stonyakin F.* Some gradient-type methods with adaptation to parameters of (δ, L)-model of objective // arXiv:1911.08425.
572. *Su W.*, *Boyd S.*, *Candes E. J.* A differential equation for modeling Nesterov's accelerated gradient method: theory and insights // JMLR. 2016. V. 17, № 153. P. 1–43.
573. *Sun H.*, *Hong M.* Distributed non-convex first-order optimization and information processing: Lower complexity bounds and rate optimal algorithms // arXiv:1804.02729.
574. *Sun Y.*, *Daneshmand A.*, *Scutari G.* Distributed Optimization Based on Gradient-tracking Revisited: Enhancing Convergence Rate via Surrogation // arXiv:1905.02637.
575. *Sutskever I. et al.* On the importance of initialization and momentum in deep learning // ICML. 2013. P. 1139–1147.
576. *Tang J. et al.* The practicality of stochastic optimization in imaging inverse problems // arXiv:1910.10100.
577. *Tarjan R. E.* Dynamic trees as search trees via euler tours, applied to the network simplex algorithm // Math. Programming. Ser. B. 1997. V. 78, № 2. P. 169–177.
578. *Taylor A.*, *Bach F.* Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions // arXiv:1902.00947.

579. *Taylor A.*, *Van Scoy B.*, *Lessard L.* Lyapunov functions for first-order methods: tight automated convergence guarantees // arXiv:1803.06073.
580. *Taylor A. B.*, *Hendrickx J. M.*, *Glineur F.* Exact worst-case convergence rates of the proximal gradient method for composite convex minimization // Journal of Optimization Theory and Applications. 2018. P. 1–22.
581. *Taylor A. B.*, *Hendrickx J. M.*, *Glineur F.* Exact worst-case performance of first-order methods for composite convex optimization // SIAM Journal on Optimization. 2017. V. 27, № 3. P. 1283–1313.
582. *Taylor A. B.*, *Hendrickx J. M.*, *Glineur F.* Smooth strongly convex interpolation and exact worst-case performance of first-order methods // Math. Programming. Ser. A. 2017. V. 161, № 1–2. P. 307–345.
583. *Titov A. et al.* Analogues of Switching Subgradient Schemes for Relatively Lipschitz-Continuous Convex Programming Problems // International Conference on Mathematical Optimization Theory and Operations Research. Cham: Springer, 2020. P. 133–149.
584. *Tran-Dinh Q.*, *Fercoq O.*, *Cevher V.* A smoothing primal-dual optimization framework for nonsmooth composite convex optimization // SIAM J. Optim. 2018. arXiv:1507.06243.
585. *Trefethen L. N.*, *Bau D.* III Numerical linear algebra. SIAM, 1997.
586. *Tropp J. A.* An introduction to matrix concentration inequalities // Foundations and Trends in Machine Learning. 2015. V. 8, № 1–2. P. 1–230.
587. *Tseng P.* On accelerated proximal gradient methods for convex-concave optimization // SIAM J. Opt. 2008 (submitted). URL: `http://www.mit.edu/~dimitrib/PTseng/papers/apgm.pdf`
588. *Tupitsa N. et al.* Alternating minimization methods for strongly convex optimization // arXiv:1911.08987.
589. *Tyrtyshnikov E. E.* How bad are Hankel matrices? // Numerische Mathematik. 1994. V. 67, № 2. P. 261–269.
590. *Uribe C. A. et al.* A dual approach for optimal algorithms in distributed optimization over networks // Optimization Methods and Software. 2020. P. 1–40.
591. *Uribe C. A. et al.* Distributed computation of Wasserstein barycenters over networks // 2018 IEEE Conference on Decision and Control (CDC). IEEE, 2018. P. 6544–6549. arXiv:1803.02933.
592. *Vaidya P. M.* A new algorithm for minimizing convex functions over convex sets // Math. Programming. Ser. A. 1996. V. 73, № 3. P. 291–341.
593. *Vaidya P. M.* Solving linear equations with symmetric diagonally dominant matrices by constructing good preconditioners // manuscript, UIUC. 1990.
594. *Vaswani S.*, *Bach F.*, *Schmidt M.* Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron // arXiv:1810.07288.
595. *Vidyasagar M.* Nonlinear systems analysis // SIAM. 2002. V. 42.
596. *Wang Y.*, *Li J.* Improved algorithms for convex-concave minimax optimization // arXiv:2006.06359.

597. *Wang Z. et al.* Cubic regularization with momentum for nonconvex optimization // arXiv:1810.03763.
598. *Wang Z. et al.* SpiderBoost and momentum: faster variance reduction algorithms // Advances in Neural Information Processing Systems. 2019. P. 2403–2413.
599. *Wang Z. et al.* SpiderBoost: a class of faster variance-reduced algorithms for nonconvex optimization // arXiv:1810.10690.
600. *Ward R., Wu X., Bottou L.* AdaGrad stepsizes: sharp convergence over nonconvex landscapes // International Conference on Machine Learning. 2019. P. 6677–6686.
601. *Weed J.* An explicit analysis of the entropic penalty in linear programming // arXiv:1806.01879.
602. *Wilson A., Mackey L., Wibisono A.* Accelerating rescaled gradient descent // arXiv:1902.08825.
603. *Wilson A. C., Recht B., Jordan M. I.* A Lyapunov analysis of momentum methods in optimization // arXiv:1611.02635.
604. *Woodworth B., Patel K. K., Srebro N.* Minibatch vs local SGD for heterogeneous distributed learning // arXiv:2006.04735.
605. *Woodworth B. E. et al.* Graph oracle models, lower bounds, and gaps for parallel stochastic optimization // Advances in Neural Information Processing Systems. 2018. P. 8505–8515.
606. *Woodworth B. E. et al.* Is local SGD better than minibarch SGD? // arXiv: 2002.07839.
607. *Woodworth B. E., Srebro N.* Tight complexity bounds for optimizing composite objectives // Advances in neural information processing systems. 2016. P. 3639–3647.
608. *Wright S.* Optimization algorithms for Data Science // IAS/Park City Mathematics Series. 2016. URL: `http://www.optimization-online.org/DB_FILE/2016/12/5748.pdf`
609. *Xiao L. et al.* Dscovr: Randomized primal-dual block coordinate algorithms for asynchronous distributed optimization // arXiv:1710.05080.
610. *Xie Y. et al.* A fast proximal point method for computing Wasserstein distance // arXiv:1802.04307.
611. *Xu Y.* Accelerated first-order primal-dual proximal methods for linearly constrained composite convex programming // arXiv:1606.09155.
612. *Xu Y.* Momentum-based variance-reduced proximal stochastic gradient method for composite nonconvex stochastic optimization // arXiv:2006.00425.
613. *Yan Y., Man X., Yang T.* Nearly optimal robust method for convex compositional problems with heavy-tailed noise // arXiv:2006.10095.
614. *Ye H. et al.* Multi-consensus decentralized accelerated gradient descent // arXiv:2005.00797.
615. *Yurtsever A., Tran-Dinh Q., Cevher V.* A universal primal-dual convex optimization framework // NIPS. 2015. P. 3150–3158.

616. *Zhang J.*, *Hong M.*, *Zhang S.* On lower iteration complexity bounds for the saddle point problems // arXiv:1912.07481.
617. *Zhang X.*, *Haskell W. B.*, *Ye Z.* A unifying framework for variance reduction algorithms for finding zeroes of monotone operators // arXiv:1906.09437.
618. *Zhou D.*, *Gu Q.* Lower bounds for smooth nonconvex finite-sum optimization // arXiv:1901.11224.
619. *Zhou D.*, *Xu P.*, *Gu Q.* Finding local minima via stochastic nested variance reduction // arXiv:1806.08782.
620. *Zhou D.*, *Xu P.*, *Gu Q.* Stochastic nested variance reduction for non-convex optimization // arXiv:1806.07811.
621. http://cvxr.com/cvx/
622. http://nbviewer.jupyter.org/github/merkulovdaniil/mipt_optimization/tree/master/; https://fmin.xyz/
623. http://www.gurobi.com/
624. https://ampl.com/
625. https://francisbach.com/chebyshev-polynomials/
626. https://github.com/amkatrutsa/MIPT-Opt
627. https://github.com/Lasagne/
628. https://github.com/oseledets/nla2018; https://nla.skoltech.ru/
629. https://github.com/Theano/
630. https://neos-server.org/neos/
631. https://pytorch.org/
632. https://simons.berkeley.edu/workshops/schedule/4791
633. https://sunju.org/research/nonconvex/#review-articles
634. https://www.di.ens.fr/~fbach/learning_theory_class/
635. https://www.gams.com/
636. https://www.ibm.com/analytics/data-science/prescriptive-analytics/cplex-optimizer
637. https://www.localsolver.com/
638. https://www.mosek.com/
639. https://www.tensorflow.org/
640. https://www.youtube.com/playlist?list=PL4_hYwCyhAvZT27DsXTISyRfQn6zo7Wud
641. https://www.youtube.com/playlist?list=PLVdkoATzj9WQM9kWYzdED3aBVW3FrnU64

Отзывы на книгу

Современные методы машинного обучения и технологии искусственного интеллекта — это, в сущности, задачи оптимизации сложных (в вычислительном смысле) функций в пространствах гигантской размерности. Их интересной особенностью является достаточно точное соответствие теоретических результатов наблюдаемым на практике эффектам. Это делает теоретическое исследование тех или иных подходов к оптимизации крайне полезным для последующего практического применения. Как правило, исследователи в области машинного обучения используют простейшие методы оптимизации, которые часто оказываются неэффективными в конкретных условиях. Для более эффективного решения задачи необходимы специфические знания, опирающиеся на современные достижения в области градиентной оптимизации. Эта книга — как раз об этом. Она может быть рекомендована всем, кто хочет научиться решать свои задачи обучения быстрее и точнее, а также тем, кто хочет внести свою лепту в создание новых методов оптимизации.

Профессор-исследователь
департамента больших данных НИУ ВШЭ,
руководитель группы байесовских методов
Дмитрий Петрович Ветров

Данная книга познакомит читателей с очень важными для практических приложений методами выпуклой оптимизации. Метод градиентного спуска и его вариации широко используются в IT-индустрии, например для решения задач машинного обучения и обработки больших данных. Телекоммуникационная компания Huawei не является исключением, методы выпуклой оптимизации применяются при разработке различных продуктов компании. В области задач беспроводной связи мы используем градиентные методы для построения и идентификации параметров моделей, описывающих передачу и усиление сигналов на базовых станциях сотовой связи.

Представленный в книге материал изложен с математической строгостью, все главы включают в себя теоретическую часть, а также примеры, задачи и упражнения, что обеспечивает читателям системный подход к изучению проблемы.

Московский исследовательский центр Huawei Co, Ltd,
технические эксперты А. Воробьёв, Е. Яницкий, О. Васюкова

Это очень интересная и своевременная книга. Автор излагает основные идеи, лежащие в основе построения и обоснования современных методов

минимизации. Включение в книгу прямодвойственных и универсальных методов делает её весьма примечательной даже на международном уровне. Заключает монографию интересный обзор современного состояния методов оптимизации, отражающий точку зрения исследователя, активно и продуктивно работающего в этой области.

Лауреат премий Данцига, фон Неймана,
ординарный профессор Католического университета Лувена
Юрий Евгеньевич Нестеров

По моему мнению, книга представляет замечательное продвинутое введение в одну из наиболее быстро развивающихся областей современной выпуклой оптимизации — методы первого порядка. Методы этого типа обладают весьма интересной теорией и имеют широчайшее поле применений, в первую очередь задачи большой размерности, где эти методы являются основным вычислительным инструментом. Будучи доступной широкой аудитории, книга в строгой и ясной форме излагает все основные разделы теории, включая наиболее современные и продвинутые её компоненты (как и следует ожидать от активно и весьма плодотворно работающего в этой области автора). Неотъемлемой и весьма ценной частью книги являются прекрасно подобранные и весьма поучительные упражнения. В целом я бы квалифицировал книгу как один из лучших известных мне учебников (да и не только учебников) по алгоритмам выпуклой оптимизации.

Лауреат премий Фалкерсона, Данцига, фон Неймана, Винера,
профессор Университета штата Джорджия (Атланта)
Аркадий Семёнович Немировский

Книга Александра Гасникова посвящена современным методам оптимизации. Автор — один из ведущих специалистов в стране в этой области. Основное внимание уделено методам выпуклой оптимизации. С одной стороны, это классическая тематика, которая присутствует во многих учебниках; с другой — это активная область современной науки, в которой каждый год получаются принципиально новые результаты как с теоретической, так и с алгоритмической точки зрения. Основная цель автора — дать это «ощущение» современной науки в области оптимизации и познакомить читателя с основными результатами. При этом используется понятный, простой язык с простыми примерами и иллюстрациями. Книгу приятно читать, и из неё можно узнать много нового и студентам, и специалистам в области математического моделирования.

Профессор СколТеха,
лауреат премии Президента РФ в области науки и инноваций
Иван Валерьевич Оселедец

Методы оптимизации — это одна из центральных дисциплин учебного плана Физтех-школы прикладной математики и информатики. И это не уди-

вительно, ведь оптимизация является важнейшим инструментом в построении и анализе различных моделей, возникающих на практике. Оптимизация лежит в основе современных методов машинного обучения и математической статистики. Даже в столь любимой мною дискретной математике оптимизация играет огромную роль при решении экстремальных задач комбинаторики и теории графов. Предлагаемая книга является весьма продвинутым пособием для углублённого изучения оптимизации, и она уже с успехом используется при проведении занятий в нашей Физтех-школе.

Директор Физтех-школы прикладной математики и информатики,
федеральный профессор
Андрей Михайлович Райгородский

Расцвет машинного обучения как части технологий искусственного интеллекта непосредственно влияет на растущий интерес к современным методам оптимизации, в особенности градиентным методам. Именно им посвящено представленное пособие, актуальность которого особенно высока для исследователей, не только применяющих имеющиеся методы, но и разрабатывающих собственные. Оно может быть полезно не только студентам, но и разработчикам. Следует отметить также интересный обзор состояния численных методов выпуклой оптимизации, представляющий самостоятельный интерес.

Ректор Университета Иннополис, профессор
Александр Геннадьевич Тормасов

Учебное издание

Гасников Александр Владимирович

Современные численные методы оптимизации.
Метод универсального градиентного спуска

Подписано в печать ??.??.2020 г. Формат $60\times90\,{}^{1}\!/_{16}$. Бумага офсетная.
Печать офсетная. Печ. л. 17. Тираж ??? экз. Заказ №

Издательство Московского центра
непрерывного математического образования
119002, Москва, Большой Власьевский пер., 11. Тел. (499) 241–08–04.

Отпечатано в ФГУП Издательство «Наука» (типография «Наука»)
121099, Москва, Шубинский пер., 6.

В соответствии с Федеральным законом №436-ФЗ
от 29 декабря 2010 года издание маркируется знаком 6+

Книги издательства МЦНМО можно приобрести
в магазине «Математическая книга»,
Москва, Большой Власьевский пер., 11. Тел. (495) 745–80–31.
E-mail: biblio@mccme.ru, http://biblio.mccme.ru