Министерство образования и науки Российской Федерации
Федеральное государственное автономное
образовательное учреждение высшего образования
«Московский физико-технический институт
(государственный университет)»

А. В. Гасников

СОВРЕМЕННЫЕ ЧИСЛЕННЫЕ МЕТОДЫ ОПТИМИЗАЦИИ.

МЕТОД УНИВЕРСАЛЬНОГО ГРАДИЕНТНОГО СПУСКА

Учебное пособие

Москва
МФТИ
2018

УДК 519.86(075)
ББК 22.17я73
Г22

Р е ц е н з е н т ы :

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*

Гасников, А. В.

Г22 Современные численные методы оптимизации. Метод универсального градиентного спуска : учебное пособие / А. В. Гасников. – М. : МФТИ, 2018. – 166 с.
ISBN 978-5-7417-0667-1

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

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

Учебное пособие является дополнительным по учебной дисциплине Оптимизация (3 курс ФУПМ).

Предназначено для студентов старших курсов, аспирантов и преподавателей МФТИ.

УДК 519.86(075)
ББК 22.17я73

Печатается по решению Редакционно-издательского совета Московского физико-технического института (*государственного университета*)

ISBN 978-5-7417-0667-1

Оглавление

Обозначения

$\mathbb{R}^n$ – n-мерное вещественное (векторное) пространство.

$\mathbb{R}^n_+ = \left\{ x \geq 0 : \ x \in \mathbb{R}^n \right\}$ – неотрицательный ортант $\mathbb{R}^n$.

const – числовая константа, значение которой зависит от контекста.

$\dim x$ – размерность вектора x, в частности, $\dim x = n$, если $x \in \mathbb{R}^n$.

$\left[x \right]_i$, x_i – i-я компонента вектора x.

$\left\| x \right\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}$ – p-норма вектора $x = \left\{ x_i \right\}_{i=1}^n \in \mathbb{R}^n$, $p \geq 1$.

$\left\langle x, y \right\rangle = \sum_{i=1}^n x_i y_i$ – скалярное произведение векторов $x, y \in \mathbb{R}^n$. Отметим, что при определении графа используется похожее обозначение $G = \left\langle V, E \right\rangle$, имеющее другой смысл.

$\left\| y \right\|_* = \sup_{\|x\| \leq 1} \left\langle x, y \right\rangle$ – сопряженная норма к норме $\| \ \|$. В частности, для p-нормы сопряженной будет q-норма, где $1/p + 1/q = 1$.

$\lceil a \rceil = \max \left\{ 1, a \right\}$.

$A \Rightarrow B$ – 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 приближенно равно числу y.

$x \sim y$ – значение выражения x пропорционально значению выражения y, например, $V = HR \Rightarrow V \sim R$ или $T = 2\pi \sqrt{l/g} \Rightarrow T \sim \sqrt{l/g}$.

$x \coloneqq y$ – x присваивается y (пришло из программирования), например, $x \coloneqq x + 1$.

$\mathrm{Lin} \left\{ z^1, .., z^m \right\}$ – линейное пространство (подпространство $\mathbb{R}^n$), «натянутое» на векторы $z^1, .., z^m \in \mathbb{R}^n$, т. е. любой элемент такого пространства можно представить в виде

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

A^T – матрица, транспонированная к матрице $A = \|A_{ij}\|_{i,j=1}^n$, т. е. $A^T = \|A_{ji}\|_{i,j=1}^n$.

I – единичная матрица, т. е. $I = \|I_{ij}\|_{i,j=1}^n$, где $I_{ij} = 0$, если $i \neq j$, $I_{ij} = 1$, иначе.

A^{-1} – матрица, обратная к квадратной матрице A, т. е. $A^{-1}A = AA^{-1} = I$.

$\mathrm{Ker}(A)^{\perp}$ – ортогональное дополнение подпространства, натянутого на собственные векторы матрицы A, отвечающие нулевому собственному значению.

$\sqrt{A}$ – квадратный корень из симметричной неотрицательно определенной матрицы A.

◊ Для каждой неотрицательно определенной симметричной матрицы существует такой ортонормированный базис, в котором действие этой матрицы можно понимать как соответствующие растяжение/сжатие/проектирование (задается собственными значениями λ_i) вдоль ортов. Тогда действие матрицы $\sqrt{A}$ можно понимать как растяжение/сжатие/проектирование (задается 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.

$$\forall\ x \in \mathbb{R}^n \to \langle x, Ax \rangle \geq 0.$$

$A \succ B$ – означает, что $A - B \succ 0$.

$1_n = \underbrace{(1,...,1)}_{n}^T$ – вектор из единиц.

$e_i = \underbrace{(0,...,0,1}_{i},0,...,0)^T$ – i-орт.

$S_n(1) = \left\{ x \in \mathbb{R}_+^n : \sum_{i=1}^n x_i = 1 \right\}$ – единичный симплекс пространства $\mathbb{R}^n$.

$B_{R,Q}(y) = \{ x \in Q : \|x - y\|_2 \leq R \}$ – пересечение евклидова шар радиуса R с центром в точке y и множества Q. Также будет встречаться множество $B_{R,Q}^{\|\,\|}(y) = \{ x \in Q : \|x - y\| \leq R \}$.

$\tilde{x} = \arg\min\limits_{x \in P} F(x)$ – означает, что $F(\tilde{x}) < F(x)$ для всех $x \in P \setminus \tilde{x}$.

$\tilde{x} \in \operatorname{Arg}\min\limits_{x \in P} F(x)$ – означает, что $F(\tilde{x}) \le F(x)$ для всех $x \in P$.

$\pi_Q(x) = \arg\min\limits_{y \in Q} \|x - y\|_2^2$ – 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. заключение).

$E_\xi\left[F(x,\xi)\right]$ – математическое ожидание по случайной величине (вектору) ξ от измеримой по ξ (вектор) функции $F(x,\xi)$. Здесь x следует понимать как параметр.

$E_\xi\left[f(\xi,\eta)\middle|\eta\right]$ – условное математическое 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.

$\lambda_{\max}(A) = \max\{\lambda : \exists\ x \ne 0 : Ax = \lambda x\}$,

$\lambda_{\min}(A) = \min\{\lambda : \exists\ x \ne 0 : Ax = \lambda x\}$.

$\sigma_{\max}(A) = \lambda_{\max}(A^T A) = \lambda_{\max}(AA^T) = \max\{\lambda : \exists\ x \ne 0 : AA^T x = \lambda x\}$.

$\tilde{\sigma}_{\min}(A) = \min\{\lambda > 0 : \exists\ x \ne 0 : AA^T x = \lambda x\}$.

$\nabla f(x) = \left(\partial f(x)/\partial x_1, ..., \partial f(x)/\partial x_n\right)^T$ – градиент гладкой функции $f(x)$, так же

$$\nabla_x f(x,y) = \left(\partial f(x,y)/\partial x_1, ..., \partial f(x,y)/\partial x_n\right)^T.$$

$\partial f(x)$ – субдифференциал выпуклой функции $f(x)$. По определению $g \in \partial f(x)$ (g – субградиент f в точке x) тогда и только тогда, когда для всех y имеет место неравенство $f(y) \ge f(x) + \langle g, y - x\rangle$.

$\nabla h(y) = \left\|\partial h_i(y)/\partial y_j\right\|_{i,j=1}^{n,m}$ – матрица Якоби гладкого отображения $h : \mathbb{R}^m \to \mathbb{R}^n$.

$\nabla^2 f(x) = \left\{\partial \nabla f(x) / \partial x_j\right\}_{j=1}^{n} = \left\|\partial^2 f(x) / \partial x_i \partial x_j\right\|_{i,j=1}^{n}$ – матрица Гессе дважды дифференцируемой функции $f(x)$. Аналогично можно определить

$$\nabla^{r+1} f(x) = \left\{\partial \nabla^r f(x) / \partial x_j\right\}_{j=1}^{n} \text{ и } \nabla^{r+1} h(y) = \left\{\partial \nabla^r h(y) / \partial y_j\right\}_{j=1}^{m}.$$

$\nabla^{r+1} f(x)[u] = \sum_{j=1}^{n} \partial \nabla^r f(x) / \partial x_j \cdot u_j$ – тензор ранга r ($u \in \mathbb{R}^n$).

$\nabla^2 h(y)[z] = \sum_{j=1}^{m} \partial \nabla h(y) / \partial y_j \cdot z_j$ – тензор ранга 2 ($z \in \mathbb{R}^m$).

$A(\text{параметры}) = \mathrm{O}\big(B(\text{параметры})\big)$ – означает, что существует такая абсолютная числовая константа C, не зависящая ни от каких параметров, что

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

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

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

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

Предисловие

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

Идея курса состояла в том, чтобы, с одной стороны, рассказать основные приемы, с помощью которых порождается многообразие современных численных методов выпуклой оптимизации первого порядка (рестарты, регуляризация, переход к двойственной задаче, адаптивная настройка на гладкость задачи 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.

Работа по подготовке данного пособия в § 1–3 была поддержана грантом РНФ 14-50-00150, в § 4 – грантом Президента РФ МД 1320.2018.1, в § 5 и приложении – грантом РНФ 17-11-01027.

Введение

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

Пожалуй, основным численным методом современной оптимизации является *метод градиентного спуска*. Метод прекрасно изложен в замечательной книге Б.Т. Поляка [61], вышедшей в 1983 году. 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 [13, 39, 61, 69, 161, 189] (решение нелинейных уравнений, поиск равновесий, обратные задачи и т. д.). Метод градиентного спуска можно использовать для задач оптимизации в бесконечномерных пространствах [250], например, для численного решения задач оптимального управления [12, 38, 39, 61, 62, 161]. 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* [31, 57, 112, 108, 119, 125, 260, 265, 271, 283].

◊ Все это также хорошо можно проследить по трем основным конференциям по анализу данных: COLT, ICML, NIPS, которые за последние 10–15 лет частично превратились в конференции, посвященные 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 [11, 36, 54, 107, 116, 119, 223, 246]. Данное пособие, подготовленное по материалам курса, прочитанного в ЛШСМ 2017, также построено по такому принципу. 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 приемов, 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 приема, не отягощая изложение техническими деталями. Градиентный спуск был

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

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 приемом 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, [61, 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 счет того, что каждая итерация станет дороже. Чтобы правильно по задаче выбрать 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. Все 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 дан-

ной. Как правило, для большинства оптимизационных задач, приходящих из практики (экономика [14, 234, 240], транспорт [15, гл. 1, 3], проектирование механических конструкций [242] и даже анализ данных [15, гл. 5], [255]), двойственная задача несет 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 сопряженную (двойственную) к ней задачу.

◊ Напомним, что при весьма общих условиях [116, гл. 5] двойственной задачей для двойственной к исходной выпуклой задаче будет исходная задача (теорема Фенхеля–Моро [47, 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 [34, п. 6.3.2], см. также правила Армихо, Вулфа, Голдстейна [11, гл. 5], [41, п. 3.1.2], [43, п. 9.4], [54, п. 1.2.3], [61, гл. 3], [246, гл. 3], выбора шага градиентного спуска. Эта концепция подготавливалась около 30 лет (см., например, [51]), и лишь весной 2013 года была оформлена Ю.Е. Нестеровым сначала в виде препринта, а потом в виде статьи [238]. 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 $\rightarrow$ гёльдерова $\rightarrow$ гладкая функция. Универсальные прямодвойственные методы, сейчас активно используются при поиске равновесий в больших транспортных сетях [5, 15]. Большая популярность самонастраивающихся оптимизационных процедур в анализе данных, особенно в глубоком обучении [31, 57, 96, 166] (в том числе использование нейросети для выбора величины шага в обучении другой нейросети), определенно указывает на то, что за адаптивными (самонастраивающимися), а по нашей терминологии «универсальными», методами будущее! Все это, безусловно, также сильно сказалось на отборе материала и сделанных в пособии акцентах.

◊ Опыт использования терминов *прямодвойственный* и *универсальный* (следуя [234, 238]) 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 [61, п. 3 § 2, гл. 8], который имеет еще 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 идет 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 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, приведенных в основном тексте пособия. Приложение написано, в первую очередь, для читателей, желающих продолжить изучение курса численных методов оптимизации. Надеемся, что приложение поможет сориентироваться читателям 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 приема (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 обсуждается *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 [52].

Чтобы приблизить изложение к «живым» лекциям и местами немного «разбавить» достаточно насыщенный формулами материал, 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 приемы (конструкции, надстройки) друг с другом, получая, таким образом, все 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 современные градиентные методы. Замечательно, что эти же структурные блоки используются и для ускоренных методов и их стохастических и рандомизированных вариантов, см. приложение, а также [13, 15, 18, 20, 24, 53, 54, 71, 88, 92, 138, 151, 156, 199, 223, 228, 238, 243].

Список литературы к пособию включает почти 300 источников, 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. – V. 8, N 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 [47, 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 методов оптимизации (выпуклой и не выпуклой). Во многом на базе именно этих двух учебников происходит обучение студентов основам численных методов оптимизации в большинстве продвинутых учебных заведениях 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 Принстонском учебнике [IV]. Этот учебник можно рекомендовать в качестве основного источника для последующего изучения.

◊ С. Бойд [113] является сейчас одним из самых цитируемых и активно публикующихся ученых 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 прежде всего о его книгах и документациях к разработанным под его руководством пакетам типа CVX [286]) являются хорошим образцом ясности изложения. Курс [I] является, пожалуй, самым известным (востребованным) в последнее десятилетие курсом по выпуклой оптимизации. ◊

В пособии имеется большое число ссылок на современную иностранную литературу. После распада Советского Союза «оптимизационный крен» сильно сместился на Запад. Однако считаем важным подчеркнуть определяющую роль российских ученых и научных школ [252] 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 еще 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 [36] 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 [8, 9, 36, 38, 39, 49, 50]. Современный учебный план студентов ФУПМ состоит из сочетания отмеченного опыта школы Н.Н. Моисеева и опыта коллег с ВМиК МГУ [11, 12, 41, 43, 67]. В данном пособии предпринята попытка посмотреть на этот учебный план, формировавшийся в течение полувека, сквозь призму современных достижений в области численных методов выпуклой оптимизации [54, 119, 223] и новых приложений [15, 31, 283]. Отметим также практикумы [287, 291] к упомянутому циклу лекций для студентов ФУПМ.

Автор также постарался учесть и обыграть в пособии некоторые наработки, которыми любезно с ним делились на всевозможных конференциях и семинарах представители различных научных школ: В.П. Булатова (Иркутск), В.Ф. Демьянова (Санкт-Петербург), И.И. Еремина (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). 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 и А.А. Шананиным.

Серьезное влияние на автора оказало регулярное общение с 2011 года 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 трем оракулам за это.

Хотелось бы отметить важную роль, которую оказала совместная научная работа, выполняемая с А.Ю. Горновым, П.Е. Двуреченским, Ф.С. Стонякиным на данный текст.

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 Е.А. Воронцовой, А.И. Голикову, М.Н. Деменкову, А.В. Чернову за ряд ценных замечаний. На ряд неточностей автору было указано учениками: Кириллом Бобыревым, Эдуардом Горбуновым, Сергеем Гуминовым, Дмитрием Камзоловым, Виктором Мишиным, Александром Рогозиным, Даниилом Селихановичем, Александром Тюриным, Ильнурой Усмановой и Салихом Хабибуллиным.

Все возможные ошибки лежат всецело на авторе. В случае обнаружения неточностей просьба присылать информацию на адрес электронной почты <gasnikov.av@mipt.ru>.

На обложке изображен спуск горнолыжников на горе Монблан (февраль 2016 г.). Примеры «горных аналогий» в оптимизации см. в [73].

§ 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, Л.В. Канторовича, Б.Т. Поляка [62]. Более современное изложение, в котором прорабатываются различные тонкие вопросы, начнется со следующего параграфа.

◊ Стоит особо отметить большой (наверное, можно даже сказать, решающий) вклад, который внес 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 [31, п. 8.7.3]; субградиентный метод Поляка [237]; *метод тяжелого 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 [61]. ◊

Рассмотрим систему обыкновенных дифференциальных уравнений

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

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

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

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

Отсюда можно сделать вывод, что любая траектория такой системы должна сходиться к *стационарной точке*[2] 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)

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

в случае достаточно малого шага h [61, гл. 2], см. также [3, 38, 52, 103, 142, 263, 272, 282]. Метод (1.3) обычно называют *методом градиентного спуска* или просто *градиентный спуск* [61], а приведенный здесь способ получения оценки скорости сходимости метода относят ко *второму методу Ляпунова* [61, § 2, гл. 2].

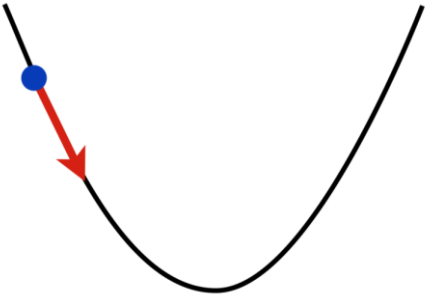

Рис. 1

Чтобы количественно оценить скорость сходимости и получить условие на выбор шага сделаем следующее предположение о *липшицевости градиента* в 2-норме [61, гл. 1]: для любых x и y имеет место неравенство

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

Из этого неравенства имеем[3] $\lambda_{\max}\left(\nabla^2 f(x)\right) \le L$, т. е. все собственные значения *матрицы Гессе* $\nabla^2 f(x) = \left\|\partial^2 f(x)/\partial x_i \partial x_j\right\|_{i,j=1}^n$ не больше L.

---

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

[3] Строго говоря, из (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. [54, п. 1.2.2].

По формуле Тейлора с остаточным членом в форме Лагранжа для любых x и y справедливо представление [68, § 58]:

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

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

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

Из неравенства (1.5) следует, что

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

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

Выбирая

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

получим

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

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

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

Из неравенства (1.7) следует: для достижения такого положения x^N, что

$$\|\nabla f(x^N)\|_2 \le \varepsilon, \tag{1.9}$$

достаточно [61, 230, 283]

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

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

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

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 как для метода вида (1.3), так и для любых других методов первого порядка, т. е. использующих только градиент функции [121, 122].

◊ 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» – существенные, см., например, [120, 121, 122, 123, 185, 230].◊

К сожалению, полученный выше результат не гарантирует сходимости даже к локальному минимуму [54, пример 1.2.2]. Впрочем, недавно было показано [207, 208], что метод (1.3) с шагом (1.6) типично сходится именно к локальному минимуму, см. также [94]. Это по-прежнему не означает сходимость к глобальному минимуму.

Если задача (1.1) является *задачей выпуклой оптимизации*, т. е. $f(x)$ – *выпуклая функция*, что означает $\lambda_{\min}\left(\nabla^2 f(x)\right)\geq 0$, то можно гарантировать сходимость метода (1.3) с шагом (1.6) к глобальному минимуму в следующем смысле [54, следствие 2.1.2]:

$$f\left(x^N\right)-f\left(x_*\right)\leq\frac{2LR^2}{N+4}, \qquad (1.11)$$

где x_* – решение задачи (1.1), $R^2=\left\|x^0-x_*\right\|_2^2$. Если решение не единственно, то под x_* в (1.11) можно понимать такое решение задачи (1.1), которое наиболее близко в 2-норме к точке старта x^0.

◊ В общем случае $f(x)$ – выпуклая функция, означает, что надграфик $f(x)$ – выпуклое множество (см. 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 [47, 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 $\partial f(x)$ [47, п. 1.5]. ◊

Из (1.7) следует, что сходимость в смысле (1.11) влечет сходимость в смысле (1.9). Рассматривая функции скалярного аргумента вида $f_M(x) = x^M$, $M \gg 1$, можно заметить, что, *сходимость по функции*, т. е. в смысле (1.11), не влечет в общем случае *сходимость по аргументу*. Точнее говоря, влечет, но скорость сходимости по аргументу может быть сколь угодно медленной.[4]

Если $f(x)$ – *μ-сильно выпуклая функция* в 2-норме, т. е. $\lambda_{\min}\left(\nabla^2 f(x)\right) \geq \mu$, $\mu > 0$, то для метода (1.3) с шагом (1.6) уже будет иметь место *линейная сходимость* (сходимость со скоростью геометрической прогрессии), причем по аргументу [119, теорема 3.10]:

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

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 условие $\lambda_{\min}\left(\nabla^2 f(x)\right) \geq \mu$ влечет (см. вывод (1.5) и [54, п. 2.1.3]) следующее условие для любых x и y:

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

Обычно условие (1.13) и понимают как определение μ-сильной выпуклой функции в 2-норме [54, п. 2.1.3].

Из неравенства (1.13) получается полезное в дальнейшем неравенство

---

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

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

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

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

Из (1.13) следует, что

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

$$= f(x) - \frac{1}{2\mu}\|\nabla f(x)\|_2^2,$$

т. е.

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

Отсюда, с учетом неравенства (1.7), имеем

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

т. е.

$$f(x^{k+1}) - f(x_*) \le \left(1-\frac{\mu}{L}\right)\left(f(x^k)-f(x_*)\right).$$

Следовательно,

$$f(x^N)-f(x_*) \le \left(1-\frac{\mu}{L}\right)^N \left(f(x^0)-f(x_*)\right) \le \exp\left(-\frac{\mu}{L}N\right)\left(f(x^0)-f(x_*)\right). \qquad (1.16)$$

Замечание 1.1 (условие градиентного доминирования). Формула (1.16) была получена в предположениях (1.4), (1.15). То есть предположение $\lambda_{\min}\left(\nabla^2 f(x)\right) \ge \mu$ на самом деле использовалось лишь в виде своего следствия (1.15). Условие (1.15) называют *условием градиентного доминирования* или *условием Поляка–Лоясиевича* [191, 239]. Приведем пример, когда это условие имеет место, однако нельзя быть уверенным даже в выпуклости $f(x)$ [42], [53, п. 4.3], [253]. Рассмотрим систему нелинейных

уравнений $g(x)=0$, записанную в векторном виде, т. е. $g:\mathbb{R}^n\to\mathbb{R}^m$, $m\le n$. Требуется найти какое-нибудь решение этой системы. Введем матрицу Якоби отображения $g:\ \partial g(x)/\partial x=\left\|\partial g_i(x)/\partial x_j\right\|_{i,j=1}^{m,n}$. Предположим, что существует такое $\mu>0$, что для всех $x\in\mathbb{R}^n$ имеет место равномерная невырожденность матрицы Якоби:

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

Тогда для функции $f(x)=\left\|g(x)\right\|_2^2$ выполняется условие (1.15) [239]. ■

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

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

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

то $W(x)=\left\|x-x_*\right\|_2^2/2$ также будет функцией Ляпунова системы (1.2). Действительно,

$$\frac{dW(x(t))}{dt}=\left\langle x(t)-x_*,\frac{dx(t)}{dt}\right\rangle=$$
$$=-\left\langle\nabla f(x(t)),x(t)-x_*\right\rangle\le f(x_*)-f(x(t))\le 0,$$
$$\frac{dW(x)}{dt}=0\ \Leftrightarrow\ x\in\operatorname{Arg}\min_{x\in\mathbb{R}^n}f(x).$$

Сделанное наблюдение «подсказывает» исследовать поведение последовательности

$$\frac{1}{2}\left\|x^k-x_*\right\|_2^2.$$

Согласно (1.3) имеем,

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

Следовательно,

$$f\left(x^k\right)-f\left(x_*\right)\overset{1}{\leq}\left\langle\nabla f\left(x^k\right),x^k-x_*\right\rangle\overset{2}{\leq}$$
$$\overset{2}{\leq}\frac{1}{2h}\left\|x^k-x_*\right\|_2^2-\frac{1}{2h}\left\|x^{k+1}-x_*\right\|_2^2+Lh\cdot\left(f\left(x^k\right)-f\left(x^{k+1}\right)\right), \tag{1.19}$$

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

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

Отсюда следует, что

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

◊ Альтернативным определением выпуклой функции является *неравенство Иенссена*: для всех x и y и произвольного $\alpha\in[0,1]$:

$$f\left(\alpha x+(1-\alpha)y\right)\leq\alpha f\left(x\right)+(1-\alpha)f\left(y\right),$$

которое по индукции влечет неравенство [116, гл. 3]:

$$f\left(\frac{1}{N}\sum_{k=1}^{N}x^k\right)\leq\frac{1}{N}\sum_{k=1}^{N}f\left(x^k\right).\ \text{◊}$$

Таким образом, ввиду выпуклости $f\left(x\right)$ имеет место неравенство

$$f\left(\overline{x}^N\right)-f\left(x_*\right)\leq\frac{LR^2}{2N}, \tag{1.20}$$

где

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

являющееся аналогом неравенства (1.11).

Резюмируем приведенные выше результаты в немного более точной и симметричной форме [138, 144, 230, 275].

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

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

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

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

Тогда

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

Если дополнительно известно, что $f(x)$ – μ*-сильно выпуклая функция в* 2*-норме, где* $\mu \ge 0$, *то*[5]

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

Приведенные в теореме 1.1 оценки скорости сходимости метода (1.22) точные. Немного могут быть улучшены только числовые множители [121, 122, 144, 275]. При получении оценки (1.23) (впрочем, как и оценки (1.10)) существенным образом использовалось, что рассматривается задача безусловной оптимизации (1.1) [230].

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

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

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

то метод имеет естественную геометрическую интерпретацию. Параболоид вращения

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

касается графика функции $f(x)$ в точке x^k и мажорирует ее на всем пространстве

---

[5] Ввиду (1.7) имеем $\min_{k=1,\dots,N} f\left(x^k\right) = f\left(x^N\right)$. 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.

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

В частности,

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

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

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

Таким образом можно прийти к основному соотношению (1.7).

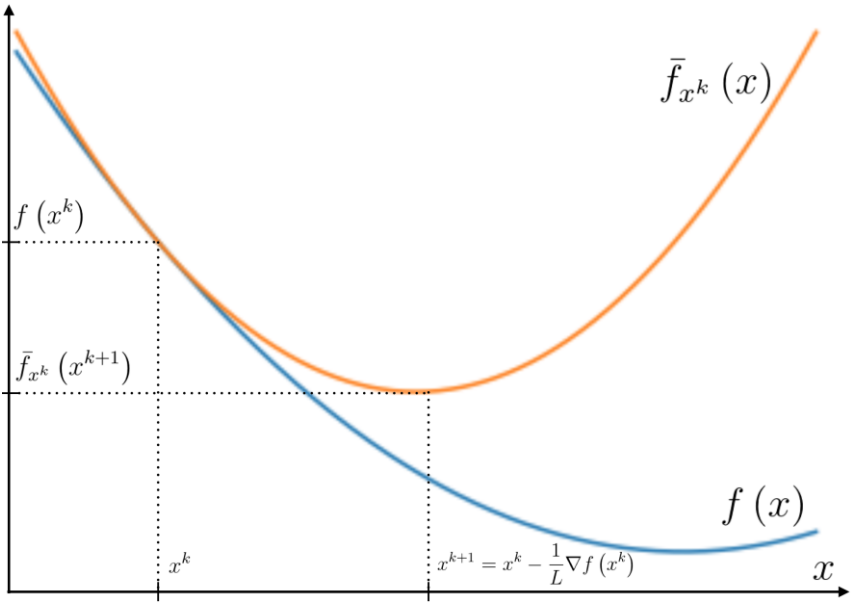

Рис. 2

Другая интерпретация имеется, например, в [61, формулы (2), (3) п. 1 § 4, гл. 1], см. также [34, п. 6.4], [41, § 5.2], [132], [246, гл. 4]. В некоторой r_k-окрестности точки x^k функция $f(x)$ заменяется линейной функцией (или более сложной моделью)

$$\tilde{f}_{x^k}(x) = f\left(x^k\right) + \left\langle \nabla f\left(x^k\right), x - x^k \right\rangle.$$

Новое положение метода определяется исходя из решения задачи

$$x^{k+1} = \arg \min_{\left\|x - x^k\right\|_2 \le r_k} \tilde{f}_{x^k}(x).$$

С помощью *принципа множителей Лагранжа* [116, гл. 5] данная задача сводится к задаче (1.25), где $L/2$ следует понимать как множитель Лагранжа к ограничению $\left\|x - x^k\right\|_2^2 \le r_k^2$. 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, [41, § 5.2]): обеспечивающего попадание исследуемого метода в область сверхлинейной (квадратичной) сходимости [34, п. 6.4], [132], [246, гл. 4]. Однако в последнее десятилетие данный подход в таком «глобализующем» контексте стал частично вытесняться *методом Ньютона с кубической регуляризацией* [227, 239] (см. также приложение), лучше изученным в теоретическом плане. Этот метод можно понимать как перезапись метода доверительной области с квадратичной моделью и ограничением вида $\left\|x - x^k\right\|_2^3 \le r_k^3$, 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. Причем в случае, когда множество Q компактно, можно выбирать, в частности, $r_k = \infty$. Получившийся в результате метод будет принадлежать к классу *методов условного градиента* [11, 61, 119, 184]. 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 [184]. 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 [1, 15, 119, 133].

Интересный взгляд на метод условного градиента предложил А.С. Немировский [72], [223, п. 5.5.3]. Оказывается, такого типа методы

можно также получать, беря 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 $1/h \equiv 0$ (в обозначениях упражнения 3.7 $1/\alpha_{k+1} \equiv 0$), и решение этой упрощенной задачи интерпретируется как приближенное решение исходной задачи (3.3). ■

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

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

где $\left\|y\right\|_* = \max\limits_{\|x\|\le 1} \langle y, x\rangle$ – сопряженная норма к норме $\|\ \|$. В этом случае неравенство (1.5) будет иметь аналогичный вид [92]:

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

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

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

Аналог неравенства (1.7) будет иметь вид [92]:

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

Отсюда можно получить следующую оценку скорости сходимости [92]:

$$f\left(x^N\right) - f\left(x_*\right) \le \frac{2L\tilde{R}^2}{N}, \tag{1.28}$$

где $\tilde{R} = \max\limits_{x:\ f(x)\le f\left(x^0\right)}\left\|x - x_*\right\|$. В случае $\|\ \| = \|\ \|_2$ оценку $\tilde{R}$ можно уточнить:

$$\tilde{R} = R = \left\|x^0 - x_*\right\|_2.$$

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

типично, что «выигрыш» в L с запасом нивелируется «проигрышем» в $\tilde{R}^2$, $\tilde{R}^2 \gg R^2$. ■

Замечание 1.4 (наискорейший спуск). Будем выбирать в методе (1.3) шаг h не из условия (1.6), а следующим образом [61, § 1, гл. 3]:

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

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:

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

наискорейший спуск сходится и не быстрее градиентного спуска с постоянным шагом [61, теорема 3 § 4, гл. 1], т. е. приведенные выше оценки скорости сходимости градиентного спуска не могут быть принципиально улучшены даже при использовании шага (1.29). Рис. 3, взятый из работы [137], демонстрирует то, как сходится метод наискорейшего спуска в этом случае.

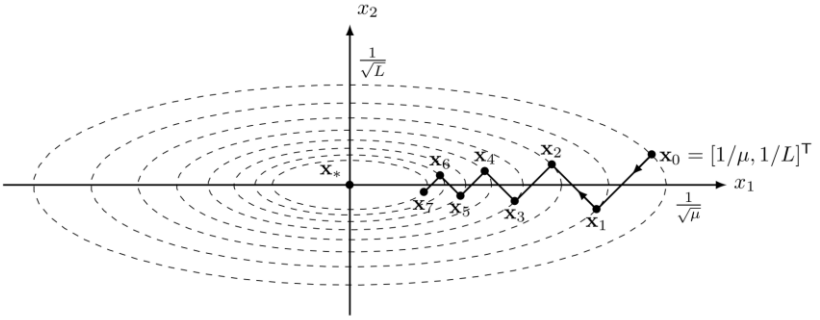

Рис. 3

Детали можно найти, например, в работе [137]. Тем не менее движение в этом направлении (использование вспомогательной одномерной / маломерной оптимизации на каждой итерации) может приносить серьезные дивиденды, см. замечания 1.5, 1.6.

Отметим также в этой связи следующий факт [61, § 1, гл. 3]. Если для минимизации положительно определенной 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 ($0<\mu=\lambda_1\le\ldots\le\lambda_n=L$), то независимо от точки старта метод будет конечен: $x^n=x_*$, где $Ax_*=b$. ■

Конструкции замечаний 1.3, 1.4, к сожалению, напрямую не переносятся на обобщения, собранные в последующих параграфах. Во всяком случае, нам о такой возможности не известно. 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 [52] – число обращений (по ходу работы метода) к оракулу за градиентом (в общем случае старшими производными – см. замечание), т. е. число обращений к подпрограмме расчета градиента, для достижения заданной точности (например, по функции) в зависимости от параметров, характеризующих класс рассматриваемых задач и желаемую точность, может быть уменьшена равномерно на всем рассматриваемом классе только на числовой множитель (в замечании 1.5 и у первого аргумента минимума в оценке (1.45) оптимальность понимается еще 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 [54, 119, 222]. Связано это с тем, что, с одной стороны, существует хорошо разработанная теория оракульной сложности задач выпуклой оптимизации [52], с другой стороны, для большинства методов первого порядка и большинства задач наиболее вычислительно затратной частью итерации является именно расчет 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 [52] стала в свое 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 (совсем 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

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

приводит к аналогичному (1.18) выражению

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

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 (1.18).

Упражнение 1.1. Докажите оценку (1.28). Почему в случае $\|\ \| = \|\ \|_2$ имеет место переход $\tilde{R}^2 \to R^2$?

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

$$x^k \in x^0 + \mathrm{Lin}\left\{\nabla f\left(x^0\right), ..., \nabla f\left(x^k\right)\right\}. \quad (1.33)$$

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

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

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

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

$$\min_{k=1,...,N} F_N\left(x^k\right) - F_N\left(x_*\right) \ge \frac{3L}{32}\frac{\left\|x^0 - x_*\right\|_2^2}{(N+1)^2},$$

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

где $F_N(x_*) = \min\limits_{x \in \mathbb{R}^n} F_N(x)$.

2) Покажите, что в этом классе методов для следующей μ-сильно выпуклой в 2-норме функции, заданной в пространстве $\mathbb{R}^\infty$ и удовлетворяющей условию (1.4) (число обусловленности $\chi = L/\mu$)

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

при всех $N \geq 1$ имеют место следующие нижние оценки:

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

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

Указание. См. [54, п. 2.1.2, 2.1.4], [61, § 2, гл. 3, п. 3 § 3, гл. 12], [119, п. 3.5], [145]. Полученные нижние оценки с точностью до мультипликативных констант достигаются на классе *ускоренных* (*быстрых*, *моментных*) *градиентных методов* [54, § 2.2]. За последние десять лет интерес к этому классу методов резко возрос, см., например, [15, 18, 24, 31, 53, 54, 72, 88, 92, 100, 102, 103, 107, 108, 119, 138, 140, 142, 143, 144, 161, 163, 175, 176, 183, 193, 194, 195, 210, 212, 219, 223, 235, 238, 248, 262, 263, 264, 272, 274, 275, 276, 279, 282, 283] и цитированную там литературу. Отметим также, что полученные нижние оценки сохраняют свой вид и для более общего по сравнению с (1.33) класса методов [52, гл. 7].

Приведем в простейшем случае основную идею ускорения градиентного спуска, следуя работам [92, 103]. Ограничимся только выпуклым случаем, отвечающим п. 1) упражнения 1.3. Про перенесение на сильно выпуклый случай см. конец § 5.

Начнем с неформальной идеи [92]. Рассмотрим два режима: 1) на текущей итерации $\left\| f(x^k) \right\|_2 \geq M$ и 2) на текущей итерации $\left\| f(x^k) \right\|_2 < M$. Каждый шаг (итерация) обычного градиентного спуска (1.22) в режиме 1 уменьшает невязку по функции согласно (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 все время пребывать в режиме 2 (в этом месте име-

ется 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 $\sim\sqrt{L/\varepsilon}$, что лучше оценки, получаемой при использовании обычного градиентного спуска $\sim L/\varepsilon$.

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

$$f\left(\overline{x}^N\right) - f\left(x_*\right) \le \frac{\sqrt{2LR^2\Delta f}}{N},$$

где

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

Следовательно, после $N_1 = \sqrt{8LR^2/\Delta f}$ итераций гарантированно бы имели $f\left(\overline{x}^{N_1}\right) - f\left(x_*\right) \le \Delta f/2$. Если *рестартовать* такую процедуру после каждого момента гарантированного уполовинивания невязки по функции, то получится метод, который достигает точности ε (по функции) после

$$\begin{aligned} N \simeq & \underbrace{\sqrt{\frac{8LR^2}{\Delta f}}}_{N_1} + \underbrace{\sqrt{\frac{8LR^2}{\Delta f/2}}}_{N_2} + \ldots + \sqrt{\frac{8LR^2}{\varepsilon}} = \\ & = \mathrm{O}\left(\sqrt{\frac{LR^2}{\varepsilon}} + \sqrt{\frac{LR^2}{2\varepsilon}} + \sqrt{\frac{LR^2}{4\varepsilon}} + \ldots\right) = \mathrm{O}\left(\sqrt{\frac{LR^2}{\varepsilon}}\right) \end{aligned} \tag{1.34}$$

итераций. Данная оценка соответствует (с точностью до числового множителя) приведенной в условии 1) упражнения 1.3 нижней оценке. Однако все 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 жестко задается в момент использования неравенства (1.7). Однако, если ввести три последовательности, связанные соотношениями ($x^0 = y^0 = z^0$),

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

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

то из (1.36)

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

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

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

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

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

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

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

если

$$x^{k+1} - z^k = Lh\cdot\left(y^k - x^{k+1}\right).$$

Получается следующая простая зависимость (*выпуклая комбинация*)

$$x^{k+1} = \tau z^k + (1-\tau)y^k, \tag{1.37}$$

где

$$\tau = \frac{1}{Lh+1}.$$

В результате для описанного здесь метода *линейного каплинга* (1.35)–(1.37) (название взято из работы [92]) можно написать следующую оценку, аналогичную (1.19),

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

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

К сожалению, описанный выше подход, во-первых, базируется на знании, как правило, априорно неизвестной величины R (используется при выборе размера шага $h=R/\sqrt{2L\Delta f}$), во-вторых, для корректности подхода под R вместо $\left\|x^0-x_*\right\|_2$ необходимо понимать заметно бóльшую величину $\max\limits_{x:\, f(x)\le f\left(x^0\right)}\left\|x-x_*\right\|$, см. также замечание 1.3. Обе отмеченные проблемы могут быть решены небольшой модификацией описанного подхода [92]. Кратко об этом написано 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 тяжелого шарика*:

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

предложенный Б.Т. Поляком в 1963–1964 гг. [61, п. 1 § 2, гл. 3]. Локальный анализ скорости его сходимости (с помощью *первого метода Ляпунова* [61, § 1, гл. 2]) при специальном выборе параметров шага $\alpha,\beta>0$ давал правильные порядки локальной скорости сходимости в сильно выпуклом случае. Однако с установлением глобальной сходимости были некоторые трудности. В частности, на специально подобранном примере с разрывным гессианом [183] метод может и не сходиться, а в чезаровском смысле траектории метода сходятся медленнее – аналогично (неускоренному) градиентному методу [162]. Несмотря на отмеченные сложности метод тяжелого шарика по-прежнему активно используется и продолжает развиваться [31, 212].

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 вспомога-

тельной маломерной оптимизации на каждой итерации, см. замечание 1.4), для которого удалось доказать глобальную сходимость с оптимальной скоростью (1.34) [55]. Метод был «забыт» почти на 20 лет. Большое внимание этот метод привлек к себе лишь после выхода в 2004 году в издательстве Kluwer на английском языке первого издания книги [54] и работы [235]. Важную роль в привлечении внимания к методу сыграла также статья [108], имеющая большое число цитирований. ◊

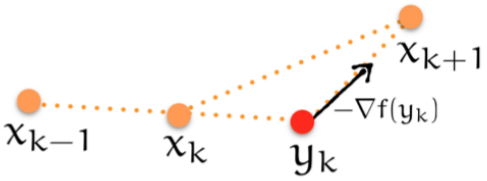

Рис. 4

Для задач безусловной минимизации наиболее популярным сейчас является следующий (двухшаговый) вариант быстрого градиентного метода [54, 272], см. рис. 4:

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

$$y^k = x^k + \frac{k-1}{k+2}\left(x^k - x^{k-1}\right),$$

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

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

Идею ускорения поясняет рис. 5, взятый из [260], на котором показаны линии уровня выпуклой функции и поведение траекторий градиентного спуска (слева) и быстрого градиентного метода (справа).

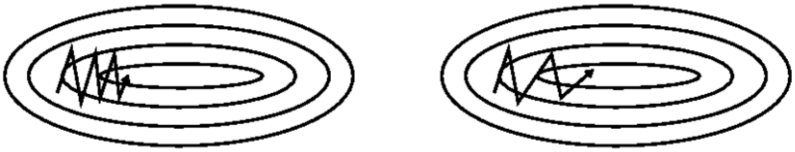

Рис. 5

Для справки приведем также вариант моментного метода для случая, когда оптимизируемая функция является μ-сильно выпуклой в 2-норме [54, (2.38) п. 2.2.1]:

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

Оба метода (1.39), (1.40) сходятся согласно соответствующим нижним оценкам, приведенным 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 1) и 2) [24], [54, теорема 2.2.2]. Однако во всех случаях (для метода (1.40) и объединенного метода) требуется априорно знать параметр μ (см. в этой связи также замечания 1.5, 1.6, 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 (1.22) и (1.39), (1.40) легко заметить, что сложность (трудоемкость) итераций у градиентного спуска и у быстрого градиентного метода практически одинаковы, 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 [143, 145, 193, 194, 195]

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

где

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

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

На классе гладких выпуклых задач метод сходится согласно оценке

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

которая достигается, например, на функции

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

Оценка (1.42) является точной оценкой скорости сходимости оптимальных методов вида (1.33) на классе гладких выпуклых задач. Другими словами, не существует такого метода вида (1.33), который бы на всем классе задач гладкой выпуклой оптимизации сходился по оценке лучшей, чем (1.42). Подчеркнем, 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

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

Такой метод также будет сходиться согласно оценке (1.42) [143].

Для класса гладких сильно выпуклых задач среди методов вида (1.33) также был найден оптимальный метод [264, 276] (с неулучшаемым числовым множителем в оценке скорости сходимости). Подобно (1.41), предложенный метод также оказался методом с конечной памятью и без вспомогательной маломерной оптимизации. Однако этот метод требует знания числа обусловленности задачи $\chi = L/\mu$, и неизвестны его адаптивные варианты по этому параметру.

Подобно (1.41) можно предложить (см. [143]) «универсальный» метод со вспомогательной трехмерной оптимизацией, который также не требует на вход никаких параметров. При этом метод равномерно опти-

мально работает на классе не только гладких выпуклых задач, но и на классе негладких задач (и задач промежуточной гладкости). В частности, для негладких задач метод сходится, согласно оценке (следует сравнить с упражнением 2.1):

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

где L_0 определяется согласно (2.4). К сожалению, непонятно, как переносить описанные выше в этом замечании методы на задачи выпуклой минимизации на множествах простой структуры.

В работе [163] был предложен такой универсальный (в смысле § 5) вариант ускоренного метода, который для невыпуклых 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) различные варианты быстрого градиентного метода начали использоваться заметно раньше [163], несмотря на невыпуклость возникающих там задач обучения нейронных сетей [31, 260, 274]. Отметим также эффективность быстрых градиентных методов в существенно невыпуклых задачах белкового фолдинга и докинга [79].

Наиболее активно ускоренные методы в последние годы исследуются в работах З. Аллена-Зу [84] и Дж. Лана [198]. ■

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

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

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 [28, 52, 54, 61, 69, 100, 246, 278], что для выпуклой задачи квадратичной оптимизации (1.30) *метод сопряженных градиентов* (первая итерация делается согласно (1.29))

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

где

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

сходится следующим образом (причем эта оценка в общем случае не может быть улучшена)

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

где $N \le n$, $\chi = L/\mu = \lambda_n/\lambda_1$, $R^2 = \left\|x^0 - x_*\right\|_2^2$ и использованы обозначения из замечания 1.4. Второй аргумент в минимуме (1.45) оценивается сверху следующим образом: $2LR^2 \exp\left(-2N/\sqrt{\chi}\right)$. Полезно сопоставить эту оценку с соответствующей (сильно выпуклой) частью оценки скорости сходимости обычного градиентного спуска (1.24): $\left(LR^2/2\right)\exp\left(-N/\chi\right)$. При $N = n$ метод гарантированно находит точное решение, что следует из последней оценки в минимуме (1.45). Сформулированный результат является фундаментальным фактом (жемчужиной) выпуклой оптимизации и вычислительной линейной алгебры одновременно, и базируется на наличии рекуррентных формул для *многочленов Чебышёва* [52, 61, 69, 100, 167, 246, 278]. Следуя [54, п. 1.3.2], приведем рассуждения, поясняющие правое равенство в (1.44). Введем обозначение (в выкладках воспользовались тем, что по условию $Ax_* = b$):

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

Заметим, что

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

т. е. Λ_k – есть *линейное подпространство Крылова*. Из определения x^{k+1} (левого равенства в (1.44)) имеем, что:

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

Введем *сопряженные направления* $\delta^k = x^{k+1} - x^k$. Сопряженные направления также порождают Λ_k:

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

Название «сопряженные направления» обусловлено свойством:

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

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

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

Из 2) и левого равенства в (1.44) (определения x^{k+1}) следует, что

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

т. е.

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

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

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

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

т. е. $\lambda_j = 0$. Таким образом, доказано правое равенство в (1.44). Оценка (1.45) получается из левого равенства (1.44) с помощью следующего наблюдения (см., например, [61, п. 2 § 2, гл. 3]): $x^N \in x^0 + \Lambda_N$ равносильно тому, что существует такой многочлен $P_N(\lambda) = 1 + a_{1N}\lambda + \ldots + a_{NN}\lambda^N$, где коэффициенты a_{1N}, …, a_{NN} могут принимать произвольные действительные значения, что $x^N - x_* = P_N(A)\left(x^0 - x_*\right)$. Поэтому для метода (1.44) должно выполняться:

$$f\left(x^N\right) - f\left(x_*\right) = \frac{1}{2}\left\langle Ax^N, x^N \right\rangle - \left\langle b, x^N \right\rangle - \underbrace{\left(\frac{1}{2}\left\langle Ax_*, x_* \right\rangle - \left\langle b, x_* \right\rangle\right)}_{0} =$$

$$= \frac{1}{2}\left\langle A\left(x^N - x_*\right), x^N - x_* \right\rangle =$$

$$= \min_{a_{1N}, \ldots, a_{NN}} \left\{ \frac{1}{2}\left\langle AP_N(A)^2\left(x^0 - x_*\right), x^0 - x_* \right\rangle \right\} \leq$$

$$\leq \frac{1}{2} \min_{P_N(\lambda):\, P_N(0)=1} \left\{ \max_{\mu = \lambda_1 \leq \lambda \leq \lambda_n = L} \left[\lambda P_N(\lambda)^2 \right] \right\}.$$

Таким образом здесь появляются многочлены Чебышёва, наименее уклоняющиеся на рассматриваемом отрезке от нуля [47, п. 6.1, гл. 2].

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).

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 – гладких сильно выпуклых задач. Конечно, хотелось бы, чтобы и для общих задач выпуклой оптимизации метод (1.44) обладал аналогичными свойствами. Однако перенести без изменений (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 (1.45) и второму (в сильно выпуклом случае), см. [52, 175, 219] 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, [61, п. 2 § 2, гл. 3], в общем случае на каждой итерации вспомогательную задачу можно решить только приближенно. В [52, § 3, гл. 7] было установлено, что достаточно решать вспомогательную задачу с относительной точностью (по функции) $\delta = \mathrm{O}\left(\varepsilon / N\left(\varepsilon\right)\right)$, где ε – желаемая относительная точность (по функции) решения исходной задачи, а $N\left(\varepsilon\right)$ – число итераций, которые делает (внешний) метод. Следовательно, вспомогательная задача может быть решена за

$$\mathrm{O}\left(\ln\left(N\left(\varepsilon\right)/\varepsilon\right)\right) = \mathrm{O}\left(\ln\left(\varepsilon^{-1}\right)\right)$$

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

$f\left(A^T x\right)+g\left(x\right)$, где вычисление $A^T x$ намного дороже по времени, чем вычисление $f\left(y\right)$ и $g\left(x\right)$, см. [175, 219] и трюк из приложения с быстрым пересчетом $A^T\left(x+\alpha \mathrm{v}\right)$ для разных $\alpha \in \mathbb{R}$:

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

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 / сильно 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 – гладких сильно выпуклых задач. Мало что известно про методы типа сопряженных градиентов (со вспомогательной маломерной оптимизацией) для задач оптимизации на множествах простой структуры [61, п. 2 § 3, гл. 7]. Также мало что известно про возможные модельные обобщения (см. § 3) и про прямодвойственность таких методов (см. § 4). Наконец, методы типа сопряженных градиентов в ряде случаев могут хуже (чем ускоренные методы с постоянными / заданными шагами) переноситься на GPU из-за вспомогательной маломерной оптимизации и, как следствие, возможности возникновения *вложенного параллелизма* [288]. С учетом того, что по теоретическим оценкам использование таких методов не позволяет в общем случае улучшать даже числовые множители (см. также замечание 1.4), то, кажется, что 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 [29], что различные варианты метода сопряженных градиентов (а их насчитывается уже, как минимум, несколько десятков [95, 160]) работают существенно быстрее ускоренных градиентных методов с постоянными шагами и их адаптивных (универсальных) вариантов. На рис. 6, взятом из работы [176], приведен характерный график сходимости одной из наиболее быстрых на практике версий метода сопряженных градиентов [29] (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 сопряженных градиентов и следующим наблюдением: методы типа сопряженных градиентов сходятся также как ускоренные методы (с фиксированными шагами) только на вырожденных (мало интересных / слишком простых для практики) примерах, см. замечание 1.4 и (1.43). В этой связи хотелось бы обратить внимание на важность проблем, затронутых в предыдущем абзаце.

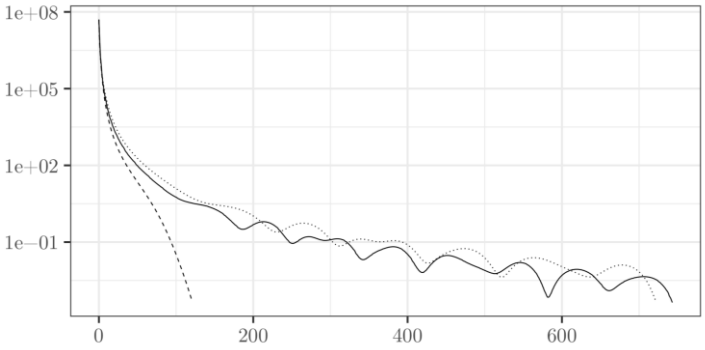

Рис. 6. По оси абсцисс – число итераций, по оси ординат – невязка по функции

Усилить сказанное в предыдущем абзаце можно следующей цитатой из уже немного устаревшей книги [25, стр. 208]: «Хотя схема сопряженных 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 похожая мысль проходит одной из основных сюжетных линий и в более современной книге [29], также посвященной 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 [248], 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 [53, п. 1.2.2], см. также упражнение 3.7. ■

◊ Наиболее популярными на практике вариантами метода сопряженных градиентов являются следующие два метода [95], [246, гл. 5]:

$$h_k = \arg\min_{h\in\mathbb{R}} f\left(x^k + hp^k\right),$$

$$x^{k+1} = x^k + h_k p^k,$$

$$p^{k+1} = \nabla f\left(x^{k+1}\right) - \beta_k p^k,\ p^0 = \nabla f\left(x^0\right),$$

$$\beta_k = -\frac{\left\|\nabla f\left(x^{k+1}\right)\right\|_2^2}{\left\|\nabla f\left(x^k\right)\right\|_2^2}, \text{ (формула Флетчера–Ривса)}$$

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

Для задач квадратичной оптимизации оба метода идентичны (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 сопряженных градиентов наиболее часто используются при решении практических задач [29] (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 $p^{k+1} = \nabla f\left(x^{k+1}\right) - \beta_k p^k$ в момент рестарта полагают $p^{k+1} = \nabla f\left(x^{k+1}\right)$. По-видимому, необходимость в таких рестартах обусловлена желанием правильно сходиться в случае оптимизации сильно выпуклых функций, см. конец § 5 и [248]. ◊

Упражнение 1.4. Предложите способы решения задач выпуклой (но не обязательно сильно выпуклой) одномерной минимизации на отрезке длины Δ за время $\mathrm{O}\left(\log\left(\Delta/\varepsilon\right)\right)$, 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 ее 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 $\mathrm{O}\left(\log\left\lceil \log\left(\Delta/\varepsilon\right)\right\rceil\right)$, 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 определенного класса методов (например, *метода центров тяжести*) размер

области (расстояние о точки старта до решения) уже не будет входить в оценку. Будет входить лишь относительная (по функции) точность. Все это верно лишь для задач на выпуклых компактах с оракулом, наделенным дополнительными нетривиальными возможностями (например, находить центр тяжести выпуклого компакта [119, 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. Причем это остается верным даже для выпуклых функций, имеющих ограниченную вариацию на полупрямой [52, упражнение 6 § 3, гл. 4]. Впрочем, в определенных случаях удается и для задач на неограниченных множествах получать оценки, в которые входит только относительная точность по функции [53, гл. 6].

Приведенная оценка $O\left(\log\left\lceil \log\left(\Delta/\varepsilon\right)\right\rceil\right)$, как оценка скорости глобальной сходимости, уже не может быть принципиально улучшена, какие бы не делались дополнительные предположения о гладкости функции и способностях локального оракула, вычисляющего значение функции и ее старшие производные в указанной точке [52, упражнение 2 § 1, гл. 8]. В частности, для класса *чебышёвских методов* [39, п. 2.9], использующих оракулы высокого порядка, можно увеличивать основание логарифмов в рассматриваемой оценке в зависимости от свойств оптимизируемой функции и порядка оракула, тем самым улучшить оценку. Однако при этом структура оценки (повторный логарифм) останется неизменной, см. приложение, а также [101, 239]. ◊

Указание. Рассмотрите, например, метод деления отрезка пополам или метод золотого сечения [11, § 3 – § 5, гл. 1]. Для анализа чувствительности методов и возможности ускорения при приближении к минимуму (в случае невырожденного минимума) стоит обратиться к [118, 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 (см. § 2) в $\mathbb{R}^n$, когда $N \geq n$, нижние оценки числа обращений к оракулу за (суб-)градиентом (что такое субградиент будет пояснено также в § 2) имеют одинаковый (с точностью до числового множителя C) вид

$$N \geq Cn\ln\left(\frac{1}{\varepsilon}\right), \tag{1.46}$$

где ε – относительная точность решения задачи по функции [52]. Оценка (1.46) достигается[6] на *методе центров тяжести* Левина–Ньюмена [52, § 3, гл. 2], [61, теорема 2 § 4, гл. 5], [119, п. 2.1]. Однако у метода центров тяжести дорогая итерация [119, п. 6.7], поэтому на практике часто используют *метод эллипсоидов* [52, § 5, гл. 2], работающий по оценке $N = \mathrm{O}\left(n^2 \ln\left(\varepsilon^{-1}\right)\right)$, но с относительно дешевой стоимостью (трудоемкости) итерации[7] $\mathrm{O}\left(n^2\right)$. 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[8] $\tilde{\mathrm{O}}\left(n^2\right)$ в смысле стоимости итерации [209]. Отметим также *метод вписанных эллипсоидов*, предложенный в 1986 г. [74, стр. 253–259], проигрывающий методу 2015 года лишь в оценке стоимости итерации.

За дополнительную мультипликативную плату, пропорциональную с точностью до логарифмических множителей размерности пространства n, написанное в предыдущем абзаце переносится и на безградиентные методы [63] – вместо (суб-)градиента оракул выдает значение функции. В гладком случае это довольно очевидное утверждение, поскольку градиент можно восстановить по частным производным, каждая из которых требует расчета значения функции в двух точках, причем 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 [74, С. 453–461]. Отметим, что на практике также часто используют и вариации метода эллипсоидов (*методы с процедурой растяжения пространства*), восходящие к работам Н.З. Шора [61, § 4, гл. 5], [65], [77]. Стоит также отметить большую роль, которые сыграли работы Н.З. Шора в появлении *субградиентного метода* (см. § 2). ◊

---

[6] С точностью до логарифмического по n множителя в случае оптимизации на евклидово-ассиметричных множествах, типа шара в 1-норме пространства $\mathbb{R}^n$. Аналогичная оговорка необходима и для выписанной далее оценки скорости сходимости метода эллипсоидов.

[7] В оценку этой стоимости изначально не входит расчет (суб-)градиента. Однако обычно (суб-)градиент можно посчитать за $\mathrm{O}\left(n^2\right)$ (см., например, начало § 2), поэтому можно считать, что выписанная оценка – есть оценка общей сложности итерации.

[8] К сожалению, с достаточно большим числовым (логарифмическим) множителем.

В случае, когда $N \le n$ (обычно это соответствует задачам оптимизации в пространстве большой размерности) и при определенных условиях в гладком сильно выпуклом случае оценка (1.46) перестает быть оптимальной (точной нижней границей сложности) [52]. Оптимальные оценки в этом случае будут достигаться на методах типа градиентного спуска, см., например, [54, 119]. ■

Упражнение 1.5 (Ю.Е. Нестеров, 2017). 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 $\sim \varepsilon/\log\left(L_0 R/\varepsilon\right)$ (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, на отрезке, высекаемом этой прямой в прямоугольнике, также с точностью $\sim \varepsilon/\log\left(L_0 R/\varepsilon\right)$ (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. Покажите, что после $\sim \log\left(L_0 R/\varepsilon\right)$ повторений такой процедуры можно найти с точностью ε (по функции) решение исходной задачи?

Упражнение 1.6 (метод условного градиента для задач квадратичной оптимизации на симплексе [15, п. 4.2.2, 4.3.3]). Рассмотрим задачу квадратичной выпуклой оптимизации:

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

где все элементы матрицы $A \succ 0$ по модулю не больше M, 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 возьмем точку старта x^0 в этой вершине. Далее действуем по индукции, шаг которой имеет следующий вид. Решаем задачу

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

Обозначим решение этой задачи через

$$y^k = (0,...,0,1,0,...,0),$$

где 1 стоит на позиции

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

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

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

Заметим, что в такой метод не входят никакие параметры!

Имеет место следующая оценка скорости сходимости описанного метода:

$$f\left(x^N\right) = f\left(x^N\right) - f_* \le \frac{2L^p R_p^2}{N},$$

где $R_p^2 = \max\limits_{x,y\in S_n(1)} \|y-x\|_p^2$, $L^p = \max\limits_{\|h\|_p \le 1} \langle h, Ah\rangle$, $1 \le p \le \infty$, причем p тут можно выбирать произвольно. С учетом того, что оптимизация происходит на симплексе, выберем $p = 1$. Несложно показать, что этот выбор оптимален. В результате получим, что $R_1^2 = 4$,

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

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

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

что может быть значительно лучше оценки

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

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).

§ 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, которые мы считаем достаточно *простым* в том смысле, что решение вспомогательной задачи проектирования на это множество (1.31) занимает существенно меньше времени, чем расчет градиента $\nabla f(x)$. В качестве наглядного примера можно рассмотреть задачу минимизации квадратичной функции на параллелепипеде, задав, например,

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

В случае плотной матрицы A расчет $\nabla f(x) = A^T \cdot (Ax)$ будет стоить $\mathrm{O}(n^2)$ *арифметических операций*[9], а проектирование на Q согласно (1.31) делается по явным формулам за $\mathrm{O}(n)$. Совсем необязательно, чтобы проектирование осуществлялось по явным формулам. Однако в подавляющем большинстве рассматриваемых в приложениях случаев простых множеств Q проектирование может быть осуществлено за (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. Тогда становится

[9] Операций типа сложения, умножения, деления двух чисел типа *float* [34, п. 1.3] – все эти операции сопоставимы (с точностью до логарифмического множителя от длины операндов) по сложности [45, глава 29], [48]. Число арифметических операций определяет время работы программы (длительность вычислений, трудоемкость), поэтому далее вместо числа арифметических операций также будет использоваться словосочетание *время работы*.

правильнее говорить уже о задаче *условной оптимизации* [61], 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 всем 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 путем специальной компактификации, вообще говоря, неограниченного множества Q, мы избавимся от отмеченной проблемы.

Другим недостатком подхода § 1 является невозможность его использования для выпуклых, но негладких функций $f(x)$. Действительно, следуя [61, § 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 § 1 субградиентом (произвольным элементом субдифференциала) в точках, в которых $f(x)$ не является гладкой.

◊ Напомним, что *субдифференциал* – это в общем случае выпуклый компакт [47, п. 1.5]. Например, для функции скалярного аргумента $f(x)=|x|$ субдифференциал будет иметь вид

$$\partial f(x)=-1,\ x<0;\ \partial f(x)=1,\ x>0 \text{ и } \partial f(x)=[-1,1],\ x=0.$$

Напомним также, что мера (Лебега) точек негладкости выпуклой функции равна нулю по теореме Радемахера [78, 218], 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 не удивительно, поскольку в это неравенство входит константа Липшица градиента L, предполагающая гладкость $f(x)$.

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

$\left|\partial f\left(x\right)\right| = 1$, 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 $\left|x^{k+1} - x^k\right| = h$, что влечет для любого k:

$$\max\left\{ f\left(x^k\right), f\left(x^{k+1}\right)\right\} \geq h/2 .$$

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 [238], мы постараемся единообразно посмотреть на гладкий и негладкий случаи.

Определим множество

$$B_{R,Q}\left(x_*\right) = \left\{x \in Q:\ \left\|x - x_*\right\|_2 \leq R\right\},$$

где x_* – решение задачи (2.1), $R = \left\|x^0 - x_*\right\|_2$. 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 $x, y \in B_{R,Q}\left(x_*\right)$

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

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

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

то (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 [140]. См. также рис. 7, соответствующий $\nu = 0$.

◊ В случае $\nu = 0$ условие (2.4) (аналогично (2.3)) выполняется для любых элементов соответствующих субдифференциалов $\partial f\left(x\right)$ и $\partial f\left(y\right)$. Фактически, условие (2.4) отвечает тому, что у функции $f\left(x\right)$ равномерно ограниченны все элементы субдифференциалов во всех точках, т. е.

функция $f(x)$ имеет равномерно ограниченную константу Липшица. Заметим, что это условие отвечает самому общему классу всех собственных выпуклых функций, поскольку требуемое свойство липшицевости должно выполняться на компакте $B_{R,Q}(x_*)$, см. также [105, 173].

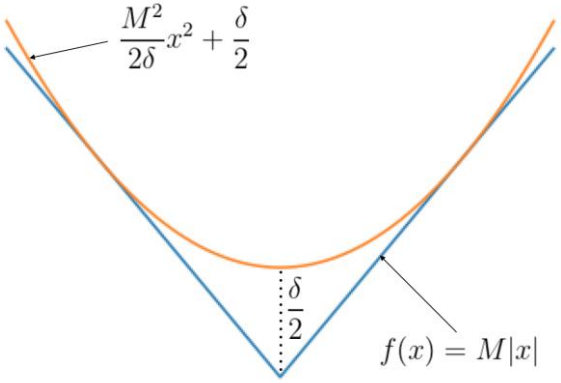

Рис. 7

Отметим также, что с точки зрения формальной логики формула (2.4) некорректна, потому что незамкнута относительно параметра ν [58]. Однако это было сделано вполне осмысленно. Дело в том, что в настоящем параграфе на формулу (2.4) будем смотреть только с точки зрения конкретного ν. В § 5 уже будем «играть» на выборе параметра $\nu \in [0,1]$, считая, что (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 противного (см. (3.4), (3.5)), использование без уточнений $\nabla f(x)$ подразумевает, что формулы справедливы при любом выборе $\nabla f(x) \in \partial f(x)$. ◊

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

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

Следствием (2.6) является условие, которое получается из (1.17) при $x = x_*$, $y = x$: для всех $x \in Q$:

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

т. е. для всех $x \in Q$:

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

Следуя [92], введем

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

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))

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

Поскольку $h \leq 1/L$, то, если $x^k, x^{k+1} \in B_{R,Q}\left(x_*\right)$, из (2.3) имеем

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

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

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

По выпуклости $f(x)$ имеем (см. (1.17)):

$$f\left(x^k\right) - f(x) \le \left\langle \nabla f\left(x^k\right), x^k - x\right\rangle, \quad (2.13)$$

также по выпуклости $f(x)$ имеем

$$f\left(\overline{x}^m\right) \le \frac{1}{m}\sum_{k=1}^{m} f\left(x^k\right), \quad (2.14)$$

где (см. (1.21))

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

Положим в (2.12) $x = x_*$, если решение не единственно, то выберем то x_*, для которого $\left\|x^0 - x_*\right\|_2^2$ минимально.

Суммируя (2.12) с учетом (2.13):

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

т. е.

$$h\cdot\left(f\left(x^{k+1}\right) - f(x_*)\right) \le h\delta + \frac{1}{2}\left\|x_* - x^k\right\|_2^2 - \frac{1}{2}\left\|x_* - x^{k+1}\right\|_2^2 \quad (2.15)$$

по $k = 0,...,m-1$, получим с учетом (2.14):

$$mh\cdot\left(f\left(\overline{x}^m\right) - f(x_*)\right) \le mh\delta + \frac{1}{2}\left\|x_* - x^0\right\|_2^2 - \frac{1}{2}\left\|x_* - x^m\right\|_2^2, \quad (2.16)$$

т. е.

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

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

$$\varepsilon = f\left(\overline{x}^N\right) - f(x_*) \approx 2\delta > \delta, \quad (2.18)$$

где

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

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

$$\frac{1}{2}\left\|x_* - x^k\right\|_2^2 \le \frac{1}{2}\left\|x_* - x^0\right\|_2^2,\ k = 0,...,N. \tag{2.19}$$

Другими словами, если $x^0 \in B_{R,Q}(x_*)$ (а это выполняется по построению $B_{R,Q}(x_*)$), то для любого $k = 0,..,N$ 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 опираясь на него самого (см. оговорку «в предположении $x^k, x^{k+1} \in B_{R,Q}(x_*)$» около формулы (2.12)). Однако несложно понять, что предполагая условие (2.3) выполненным и вне множества $B_{R,Q}(x_*)$, то есть на всем 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 $\left\{x^k\right\}_{k=0}^{N}$ никогда не выйдет за пределы множества $B_{R,Q}(x_*)$ и поэтому от того, что именно мы предполагали о выпуклой на всем множестве Q функции $f(x)$ вне множества $B_{R,Q}(x_*)$, ничего не зависит – мы никогда не окажемся вне $B_{R,Q}(x_*)$.

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

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

Вспоминая, что на h было условие $h \le 1/L$, аналогично (1.6) выберем

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

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

$$f\left(\overline{x}^N\right) - f(x_*) \le \frac{LR^2}{2N} + \delta. \tag{2.22}$$

Замечание 2.1 (условие слабой квази-выпуклости). Вместо $\overline{x}^N$ в приведенных выше формулах можно использовать

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

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

$$\left(f\left(x^k\right) - f\left(x_*\right)\right) \le \left\langle \nabla f\left(x^k\right), x^k - x_* \right\rangle,$$

т. е. «полноценная» выпуклость $f(x)$ не требуется.

Отметим, что условие (2.13) также ослабляют следующим образом (следует сравнить с однородными относительно x_* функциями [61, п. 4 § 3, гл. 3] и звездной выпуклостью [239]):

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

Условие (2.24) иногда называют *условием α-слабой квази-выпуклости* функции $f(x)$. В последнее время оно стало достаточно популярно в связи с приложениями, возникающими в *Глубоком Обучении* (*Deep Learning*) [179]. Несложно показать, что приведенные выше рассуждения переносятся и на этот случай. При этом оценка (2.22) «портится» следующим образом [175]:

$$f\left(\hat{x}^N\right) - f\left(x_*\right) \le \frac{LR^2}{2\alpha N} + \delta,$$

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

В приложениях часто используют также такое неравенство

$$\begin{aligned} f\left(\overline{x}^N\right) - f\left(x_*\right) &\le \underbrace{\sup_{x \in Q} \frac{1}{N} \sum_{k=0}^{N-1} \left\langle \nabla f\left(x^k\right), x^k - x \right\rangle}_{\text{сертификат точности}} \le \\ &\le \frac{f\left(x^0\right) - f\left(x^N\right) + L \sup_{x \in Q} \left\| x - x^0 \right\|_2^2}{2N} + \delta, \end{aligned} \tag{2.25}$$

где (следует сопоставить с (1.21))

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

Введенный в (2.25) *сертификат точности* (accuracy certificate) играет ключевую роль в обосновании прямодвойственности исследуемого мето-

да [226] (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):

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

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

$$\|\nabla f(y) - \nabla f(x)\|_* \le L_\nu \|y - x\|^\nu, \ \nu \in [0,1], \ L_0 < \infty. \tag{2.27}$$

Тогда (2.26) имеет место с константой L, рассчитываемой по формуле (2.5). Попробуем, следуя А.С. Немировскому [223], распространить метод градиентного спуска на этот случай. Как уже отмечалось в предыдущем параграфе, сделать это за счет такого обобщения метода не получается (см. формулу (1.27)):

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

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

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

Получим условия на выпуклую по x функцию $V(x, x^k)$, при которых приведенная в § 2 *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

$$\begin{aligned} \langle h \nabla f(x^k), x^{k+1} - x \rangle &\le \langle \nabla_{x^{k+1}} V(x^{k+1}, x^k), x - x^{k+1} \rangle \overset{1}{=} \\ &\overset{1}{=} V(x, x^k) - V(x, x^{k+1}) - V(x^{k+1}, x^k). \end{aligned} \tag{2.30}$$

Таким образом, достаточно потребовать выполнение равенства 1[10] тождественно по x. Например, если считать, что имеет место следующее представление:

$$V(x,y)=d(x)-d(y)-\langle\nabla d(y),x-y\rangle \tag{2.31}$$

с выпуклой функцией $d(x)$, то тождество 1 также имеет место. Вторым и заключительным ключевым местом было неравенство (2.11), которое в нашем случае останется верным, если

$$V\left(x^{k+1},x^k\right)\geq\frac{1}{2}\left\|x^{k+1}-x^k\right\|^2. \tag{2.32}$$

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 $V(x,y)$ – порожденным ею *расхождением* или *дивергенцией Брэгмана* (Bregman divergence) [10, 119, 223]. Отметим, что для метода (2.29) в оценку (2.22) будет входить $R^2=2V\left(x_*,x^0\right)$. Если решение не единственно, то оценка (2.22) будет верна в том числе и для того решения x_*, которое доставляет минимум R^2. Отметим также, что для «сохранения конструкции» достаточно, чтобы условия (2.26), (2.27) имели место только при

$$x,y\in B_{R,Q}^{\|\ \|}(y)=\left\{x\in Q:\ \|x-y\|\leq R\right\}.$$

Рассуждения здесь аналогичны рассуждениям, использованным при выводе (2.19).

Примеры прокс-функций для множеств Q вида шаров в различных нормах собраны в табл. 1. Параметр

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

Дополнительно к тому, что приведено в табл. 1, особо отметим «Spectrahedron setup» [119, п. 4.3]. Приведенные в табл. 1 прокс-функции можно распространить и на прямые произведения шаров [223, п. 5.3.3].

---

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

Таблица 1

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

По-видимому, в табл. 1 в общем случае нельзя избавиться от дополнительного $\ln n$ фактора (множителя) 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 работе в евклидовом случае, 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

$$d(x) = \sum_{i=1}^n x_i \ln x_i\ ,\ V(x,y) = \sum_{i=1}^n x_i \ln\left(x_i / y_i\right)\ \text{ и }\ x^0 = \left(n^{-1},...,n^{-1}\right).$$

Все приведенные в табл. 1 прокс-функции 1-сильно выпуклы в указанных нормах на всем пространстве. Поэтому их можно использовать и в том случае, когда мы заранее не знаем, где локализовано решение [1]. Например, если стартовать с $x^0 = 0$ 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 (2.22) от выбора нормы не будет сильно зависеть, в то время как для константы

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

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

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

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

Как мы увидим в дальнейшем (см. упражнение 4.6), задача (2.29) при известном векторе $\nabla f\left(x^k\right)$ для примеров из табл. 1 решается за время $\mathrm{O}\left(n \ln^2\left(n/\varepsilon\right)\right)$, где ε – точность решения в смысле сходимости по аргументу. Выделим особо один частный случай, когда эту оценку можно улучшить до $\mathrm{O}(n)$:

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

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

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

Отметим, что метод (2.29) при $h \le 1/L$ ввиду $V(x,y) \ge \|x-y\|^2/2$ имеет геометрическую интерпретацию, аналогичную обычному градиентному спуску (см. замечания 1.2, 1.3).

Конструкция (2.7) – (2.12) с учетом (2.29), (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 (2.4) с $\nu = 0$ оценка (2.22) для метода (2.6), (2.21) с

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

т. е. с[11]

[11] То, что для негладких задач шаг градиентного метода $h = \mathrm{const} \cdot \varepsilon/L_0^2$, а для гладких – $h = \mathrm{const}/L_1$, можно было понять и из П-теоремы теории размерностей [2, 37].

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

будет иметь вид

$$f\left(\overline{x}^N\right) - f\left(x_*\right) \le \frac{L_0 R}{\sqrt{N}}. \tag{2.35}$$

Покажите, что в классе методов

$$x^{k+1} \in x^0 + \mathrm{Lin}\left\{\nabla f\left(x^0\right), ..., \nabla f\left(x^k\right)\right\} \tag{2.36}$$

оценка (2.35) при $N \le n$, где $n = \dim x$, не может быть улучшена с точностью до замены в знаменателе в правой части $\sqrt{N} \to \sqrt{N+1}$, см. замечание 1.5.

Указание. Следует сравнить это упражнение с упражнением 1.3. Согласно (2.5) $L = L_0^2 \big/ (2\delta)$. С учетом этого найдите минимум правой части неравенства (2.22) по $\delta > 0$. Получите отсюда оценку (2.35). Для получения нижней оценки воспользуйтесь, например, [54, п. 3.2.1], [119, п. 3.5]. По заданному $N \le n$ определите

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

Из решения задачи

$$L_0\tau + \frac{\mu N}{2}\tau^2 \to \min_{\tau}$$

определите $\tau_* = -R\big/\sqrt{N}$. Тогда

$$\|x_*\|_2^2 = N\tau_*^2 = R^2, \quad f\left(x_*\right) = \min_{x \in \mathbb{R}^n} F_N(x) = -L_0 R \Big/ \left(2\sqrt{N}\right) = F_N^*.$$

Если $x^0 = 0$, тогда для метода (2.36) после N итераций имеет место: $x_i^N = 0$ при $i > N$. Таким образом,

$$F_N\left(x^N\right) - F_N^* \ge -F_N^* = \frac{L_0 R}{2\sqrt{N}} = \frac{L_0^2}{2\mu N}. \tag{2.37}$$

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 [52, гл. 4]. ■

Упражнение 2.2. Покажите, что при $\delta > 0$ оценку (2.22) нельзя принципиально улучшить в части зависимости от N, не ухудшив в части зависимости ее от $\delta > 0$.

Указание. Допустите противное, т. е. что можно получить следующую оценку:

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

Рассмотрите негладкий случай, т. е. используйте (2.4) с $\nu=0$. Согласно нижним оценкам (см. упражнение 2.1), для любого $N\le n$, и для любого метода вида (2.36) существует такая выпуклая функция из гёльдерова класса с $\nu=0$ и константой L_0, что

$$f\left(\bar{x}^N\right)-f\left(x_*\right)\ge\frac{L_0R}{2\sqrt{N}}\,.$$

Покажите, что при достаточно большом N (а следовательно, и n) это противоречит (2.38). Для этого, согласно (2.5), подставьте в (2.38) $L=L_0^2/\left(2\delta\right)$ и специально подберите $\delta=L_0R\sqrt{C_1/\left(2C_2N^{1+\gamma}\right)}$. Тогда

$$f\left(\bar{x}^N\right)-f\left(x_*\right)\le C_1\frac{L_0^2R^2}{2\delta N^{1+\gamma}}+C_2\delta=\sqrt{2C_1C_2}\,\frac{L_0R}{\sqrt{N^{1+\gamma}}}\,.\ \blacksquare$$

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

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

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

$$f\left(\tilde{x}^N\right)-f\left(x_*\right)\le\frac{512L_0^2}{\mu N}+2\delta,$$

точнее, по оценке

$$f\left(\tilde{x}^N\right)-f\left(x_*\right)\le\frac{512L_0^2}{\mu N},\ \ N\le\frac{256L_0^2}{\mu\delta}? \tag{2.39}$$

◊ При $\delta=0$ существует много способов уменьшения константы 512 в оценке (2.39) на два порядка [181, 186, 188, 197, 254]. 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 рестарты получают оптимальный метод для сильно выпуклых задач*. Во всяком случае пока не удалось придумать

контрпримера, 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, с возможно 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 ее обоснования) существенно завязано именно на рестарты по расстоянию от текущей точки до решения (или невязке по функции). Использование более удобных критериев для рестарта, например для гладких выпуклых задач безусловной оптимизации можно было бы использовать легко вычислимую величину нормы градиента функционала, к сожалению, не позволяет строго обосновать сохранение свойства оптимальности метода в общем случае [248], см. также замечание 5.3. ◊

2) Попробуйте обобщить этот результат на случай, когда используется произвольная норма $\|\ \|$ и не евклидова прокс-структура, однако имеет место следующее свойство: $d(x) \le C_n \|x\|^2$. Покажите, что при таком предположении оценка (2.39) немного ухудшится: $\mu \to \mu / C_n$. Заметим, что для прокс-функций из табл. 1 $C_n = \mathrm{O}(\ln n)$.[12]

Указание. 1) См., например, [51, 52, 186, 188, 222, 229, 259]. Покажите, что при $N_1 \le L_0^2 R_0^2 / \delta^2$

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

Неравенство 1 имеет место ввиду μ-сильной выпуклости $f(x)$ (см. (1.14)), неравенство 2 – ввиду (2.35), а неравенство 3 – ввиду $N_1 \le L_0^2 R_0^2 / \delta^2$. Выберите N_1 из условия $R_1 = R_0 / 2$. Из (2.40) получите $N_1 \simeq 64 L_0^2 / \left(\mu^2 R_1^2\right)$.

◊ 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,

[12] 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.

либо просто неизвестен, либо грубо оценен, 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 лучше, чем рестарты по этому неизвестному параметру [229, 246]: решаем задачу с $\mu = \mu_0$, метод не сходится, 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 [16]. Впрочем, в некоторых случаях можно более изящно решать отмеченную проблему [157, 159, 193, 248, 257, 259, 272]. Мы вернемся к этому вопросу в конце § 5. ◊

Далее, после N_1 итераций, рестартуйте исходный метод и положите $x^0 := \overline{x}^{N_1}$. Определите $N_2 \le L_0^2 R_1^2 / \delta^2$ из условия $R_2 = \left\| \overline{x}^{N_2} - x_* \right\|_2 = R_1/2$. Получите, что $N_2 \simeq 64 L_0^2 / \left(\mu^2 R_2^2 \right)$ и т. д. … После k таких рестартов общее число итераций будет

$$N = N_1 + ... + N_k \simeq \frac{256 L_0^2}{\mu^2 R_0^2} \left(1 + 4^1 + ... + 4^k \right) \approx \frac{4^{k+5} L_0^2}{\mu^2 R_0^2}. \qquad (2.41)$$

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

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

Покажите, что из $N_k \le L_0^2 R_{k-1}^2 / \delta^2$, с учетом (2.42), следует $\varepsilon \ge 2\delta$. Покажите, что из (2.41) следует

$$\frac{\mu R_k^2}{2} = \frac{512 L_0^2}{\mu N}. \qquad (2.43)$$

Объединяя (2.42), (2.43) и $\varepsilon \ge 2\delta$, получите оценку (2.39). ■

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

Указание. См. [223, п. 5.6]. ■

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

Указание. См. [23]. ■

§ 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 имеем (δ, L)-*модель функции* $f(x)$ в точке x (относительно нормы $\|\ \|$), и обозначать эту модель $\left(f_\delta(x); \psi_\delta(y,x)\right)$, если для любого $y \in Q$ справедливо неравенство [72]:

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

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

◊ Из (3.1) при $y = x$ следует, что $0 \le f(x) - f_\delta(x) \le \delta$, 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 $y \in Q$

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

Частным случаем, отвечающим условиям

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

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 [140] , см. также [61, гл. 4], [102, 136, 268]. Близкие концепции модели функции также имеются в работах [215, 247].

◊ 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 (3.1) интерпретируют уже не как условие гладкости $f(x)$ (липшицевости

градиента), а как *условие относительной гладкости* [104, 213]. Важное преимущество, которое приобретается в случае такой замены, – отсутствие условия (2.32) на дивергенцию Брэгмана $V(y,x)$ (правда, и некоторые сложности приобретаются, например, задача (3.21) может стать более сложной в таком случае). 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

$$x,y\in B^{\|\|}_{R,Q}(y)=\left\{x\in Q:\ \|x-y\|\le R\right\},$$

где $R^2=V\left(x_*,x^0\right)$, см. вторую половину § 2. Если решение не единственно, то в определении R^2 выбирается такое решения x_*, которое доставляет минимум R^2. ◊

Заметим, что (3.1) включает в себя намного больше свободы (см. [140]) по сравнению с (2.3). В частности, (3.1) включает возможность неточного вычисления (суб-)градиента и значения функции, а не только игру на гладкости (см. (2.4), (2.5)). Мы вернемся к более подробному обсуждению вопросов, связанных с концепцией (3.1), ниже, см. примеры 1, 2 § 3 и упражнения 3.2, 3.3, 4.3.

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

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

где $V\left(x,x^k\right)$ – дивергенция Брэгмана, определенная в конце предыдущего параграфа. Если задача (3.3) точно решена, то существует такой

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

что для любого $x\in Q$

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

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

$$\left\langle \nabla_{x^{k+1}} \Psi\left(x^{k+1}, x^k\right), x_* - x^{k+1} \right\rangle \geq -\tilde{\delta},$$

т. е. (следует сравнить с [102] и [223, п. 5.5.1.2])

$$\left\langle \nabla_{x^{k+1}} \Psi\left(x^{k+1}, x^k\right), x^{k+1} - x_* \right\rangle \leq \tilde{\delta}, \tag{3.4}$$

где $\tilde{\delta} > 0$. Добиться выполнения (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 $x = x_*$ (3.4). В этом случае нужно предполагать, что существует такой

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

что

$$\max_{x \in Q} \left\langle \nabla_{x^{k+1}} \Psi\left(x^{k+1}, x^k\right), x^{k+1} - x \right\rangle \leq \tilde{\delta}. \tag{3.5}$$

Введем $\mathrm{Prog}^h_{\psi, V}\left(x^k\right)$, см. также (2.9), (2.11):

$$\mathrm{Prog}^h_{\psi, V}\left(x^k\right) = -\left(\psi_\delta\left(x^{k+1}, x^k\right) + \frac{1}{h} V\left(x^{k+1}, x^k\right)\right). \tag{3.6}$$

Из выпуклости $\psi_\delta\left(x, x^k\right)$ по x, определения x^{k+1} (формула (3.3)) и тождества 1 в (2.30), подобно выводу (2.10), из (3.5) (или из (3.4), в этом случае можно сразу положить в последующих выкладках $x = x_*$) получим

$$\begin{aligned} -\tilde{\delta} \leq \left\langle \nabla_{x^{k+1}} \Psi\left(x^{k+1}, x^k\right), x - x^{k+1} \right\rangle &= \left\langle \nabla_{x^{k+1}} \psi_\delta\left(x^{k+1}, x^k\right) + \frac{1}{h} \nabla_{x^{k+1}} V\left(x^{k+1}, x^k\right), x - x^{k+1} \right\rangle = \\ = \left\langle \nabla_{x^{k+1}} \psi_\delta\left(x^{k+1}, x^k\right), x - x^{k+1} \right\rangle &+ \frac{1}{h} V\left(x, x^k\right) - \frac{1}{h} V\left(x, x^{k+1}\right) - \frac{1}{h} V\left(x^{k+1}, x^k\right) \leq \\ \leq \psi_\delta\left(x, x^k\right) - \psi_\delta\left(x^{k+1}, x^k\right) &+ \frac{1}{h} V\left(x, x^k\right) - \frac{1}{h} V\left(x, x^{k+1}\right) - \frac{1}{h} V\left(x^{k+1}, x^k\right). \end{aligned} \tag{3.7}$$

Отсюда следует, что

$$-\psi_\delta\left(x, x^k\right) \leq \mathrm{Prog}^h_{\psi, V}\left(x^k\right) + \tilde{\delta} + \frac{1}{h} V\left(x, x^k\right) - \frac{1}{h} V\left(x, x^{k+1}\right). \tag{3.8}$$

Согласно неравенству (3.1) при $y = x = x^k$:

$$0 \leq f\left(x^k\right) - f_\delta\left(x^k\right) \leq \delta. \tag{3.9}$$

Отсюда по левому неравенству (3.1) при $y = x$, $x = x^k$:

$$f\left(x^k\right) - f\left(x\right) - \delta \le f_\delta\left(x^k\right) - f\left(x\right) \le -\psi_\delta\left(x, x^k\right). \tag{3.10}$$

При $h \le 1/L$ из (3.6) имеем

$$\begin{aligned} &\mathrm{Prog}^h_{\psi, V}\left(x^k\right) = -\left(\psi_\delta\left(x^{k+1}, x^k\right) + \frac{1}{h} V\left(x^{k+1}, x^k\right)\right) \overset{1}{\le} \\ &\overset{1}{\le} f_\delta\left(x^k\right) - f\left(x^{k+1}\right) + \delta \overset{2}{\le} f\left(x^k\right) - f\left(x^{k+1}\right) + \delta. \end{aligned} \tag{3.11}$$

Неравенство 1 следует из (2.32) и правого неравенства (3.1) при $x = x^k$, $y = x^{k+1}$, неравенство 2 следует из (3.9). Подставляя неравенство (3.11) в (3.8), получим

$$-\psi_\delta\left(x, x^k\right) \le f\left(x^k\right) - f\left(x^{k+1}\right) + \tilde{\delta} + \delta + \frac{1}{h} V\left(x, x^k\right) - \frac{1}{h} V\left(x, x^{k+1}\right). \tag{3.12}$$

Подставляя (3.10) в (3.12), получим аналог неравенства (2.15):

$$f\left(x^k\right) - f\left(x\right) \le f\left(x^k\right) - f\left(x^{k+1}\right) + \tilde{\delta} + 2\delta + \frac{1}{h} V\left(x, x^k\right) - \frac{1}{h} V\left(x, x^{k+1}\right),$$

т. е. при $h \le 1/L$ имеет место основное неравенство

$$f\left(x^{k+1}\right) - f\left(x\right) \le \frac{1}{h} V\left(x, x^k\right) - \frac{1}{h} V\left(x, x^{k+1}\right) + \tilde{\delta} + 2\delta. \tag{3.13}$$

Мы остановимся на этой формуле, поскольку все дальнейшие рассуждения в точности совпадают с аналогичными рассуждениями из предыдущего параграфа. Общий вывод, который можно сделать из (3.13), сформулируем следующим образом.

Теорема 3.1. *Пусть нужно решить задачу* (2.1). *Для метода* (3.3), (2.21):

$$x^{k+1} = \arg_{\tilde{\delta}} \min_{x \in Q} \left\{\psi_\delta\left(x, x^k\right) + LV\left(x, x^k\right)\right\}, \tag{3.14}$$

в условиях (3.1), (3.4) *имеют место оценки, аналогичные оценкам*[13] (2.22), (2.19):

$$f\left(\overline{x}^N\right) - f\left(x_*\right) \le \frac{LR^2}{N} + \tilde{\delta} + 2\delta, \tag{3.15}$$

где

[13] С оговоркой, аналогичной (2.18), в случае оценки (3.16).

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

$$V\left(x_*, x^k\right) \le V\left(x_*, x^0\right), \tag{3.16}$$

$R^2 = V\left(x_*, x^0\right)$. *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* [53], демонстрирующих полезность рассмотрения более общих ситуаций, чем (3.2).

Пример 3.1 (композитная оптимизация). Рассмотрим задачу *композитной оптимизации* (composite optimization) [108, 229]:

$$f(x) = F(x) + g(x) \to \min_{x \in Q} \tag{3.17}$$

с выпуклой функцией $F(x)$, удовлетворяющей условию (2.3), и, вообще говоря, негладкой выпуклой функцией $g(x)$ простой структуры. Последнее означает, что множества Лебега

$$\Lambda_y = \left\{x \in Q:\ g(x) < y\right\} \tag{3.18}$$

имеют простую структуру. К такой задаче, например, можно отнести задачу LASSO:

$$\frac{1}{2}\left\|Ax - b\right\|_2^2 + \lambda\left\|x\right\|_1 \to \min_{x \in \mathbb{R}^n}.$$

Естественным обобщением метода (2.29) для задачи (3.17) будет

$$x^{k+1} = \arg\min_{x \in Q}\left\{\left\langle \nabla F\left(x^k\right), x - x^k\right\rangle + g(x) + LV\left(x, x^k\right)\right\}. \tag{3.19}$$

Метод (3.19) в точности соответствует методу (3.14) с

$$\psi_\delta(y, x) = \left\langle \nabla F(x), y - x\right\rangle + g(y) - g(x). \tag{3.20}$$

Таким образом, все приведенные выше результаты удается 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 все становится на места, если заметить, что по принципу множителей Лагранжа [47, п. 2.1], использованном в «обратном направлении», при весь-

ма общих условиях существует такое y, что задача (3.17) эквивалентна задаче

$$F(x) \to \min_{x \in \Lambda_y}$$

с множеством Λ_y (см. (3.18)) простой структуры.[14]

Беря в качестве композитного члена индикаторные функции выпуклых множеств простой структуры, можно получить результаты § 2 из композитного подхода с $Q = \mathbb{R}^n$.

Беря в качестве композитного члена линейные функции, несложно понять, что скорость сходимости метода (3.19) в негладком случае (см. (2.3) – (2.5) с $\nu = 0$) зависит от константы L_0, а не от константы Липшица оптимизируемого функционала [108, 233]. Это можно понять непосредственно из самой оценки (2.22), т. е. без композитной оптимизации. Однако с композитной оптимизацией это свойство становится более ясным. ■

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 $\delta = 0$)

$$\psi_\delta(y, x) = f(y) - f(x)$$

и выбрать произвольное L в условии (3.1), то полученный по формуле (3.3) метод

$$x^{k+1} = \arg\min_{x \in Q} \left\{ f(x) + LV\left(x, x^k\right) \right\} \tag{3.21}$$

становится известным *прокс-методом* решения задачи (2.1) [61, § 1, гл. 6], [114], [128]. Метод (3.21) будет сходиться согласно оценке (3.15) из теоремы 3.1, т. е. быстрее, чем следует 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 $LV\left(x, x^k\right)$, будет не меньше, чем $\tilde{O}\left(L_0^2 \big/ \left(L\tilde{\delta}\right)\right)$, где L_0 опре-

[14] Есть и другой способ, объясняющий факт отсутствия в оценке скорости сходимости информации о $g(x)$ [18, замечание 6].

деляется по формуле (2.4).[15] 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 $\mathrm{O}\left(L_0^2 \big/ \left(L\tilde{\delta}\right)\right)$, выбирая $\tilde{\delta} \sim \varepsilon$, где ε – желаемая точность (по функции) решения исходной задачи (2.1), получим оценку вида (2.35), что уже соответствует нижней оценке (2.37). Действительно, выбирая N в (3.15) из условия $LR^2/N \sim \varepsilon$, получим для итоговой сложности

$$N \frac{L_0^2}{L\tilde{\delta}} \sim \frac{LR^2}{\varepsilon} \frac{L_0^2}{L\varepsilon} = \frac{L_0^2 R^2}{\varepsilon^2}, \tag{3.22}$$

что соответствует оценке (2.35), если приравнять правую часть (3.22) N из (2.35) и выразить $\varepsilon(N)$.

◊ На первый взгляд, кажется, что нет никакой выгоды от описанного в предыдущем абзаце подхода. Однако выгода получается в случае, когда приведенную 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 $V(x,y) = \frac{1}{2}\|x-y\|_2^2$ можно показать, что исходную задачу (3.17) можно решить с точностью по функции ε за $\mathrm{O}\left(L_{0,g}^2 R^2 / \varepsilon^2\right)$ обращений к оракулу за субградиентом $g(x)$, где $L_{0,g}$ определяется согласно (2.4) с $\nu = 0$ и $f \equiv g$, и $\mathrm{O}\left(L_{1,F} R^2 / \varepsilon\right)$ обращений к оракулу за градиентом $F(x)$, где $L_{1,F}$ определяется согласно (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 $F(x)$ занимает намного больше времени, чем вычисление субградиента $g(x)$, такое расщепление дает очевидные

[15] 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 [15, 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».

преимущества [187, 200]. Прием, с помощью которого удается достичь описанного результата, называется *градиентный слайдинг* [200]. По-видимому, впервые он был предложен в работе [187]. В последние годы этот прием стал достаточно популярным в связи с большим числом приложений в задачах анализа изображений. За многочисленными обобщениями и приложениями данного приема можно следить, например, по работам Дж. Лана [198].

К сожалению, техника слайдинга требует довольно тонких и весьма громоздких рассуждений для своего обоснования, см., например, упражнение 3.3. Пожалуй, это единственная известная нам и достаточно широко использующаяся конструкция в современной выпуклой оптимизации, суть которой пока так и не удалось раскрыть с помощью элементарных соображений. Для всех остальных основных конструкций в данном пособии предпринимается попытка представить их достаточно простым (естественным) способом. ◊

Заметим также, что в работах [146, 210, 211] на базе *проксимального подхода* (описанного выше, см. (3.21)) был предложен новый общий способ ускорения различных неускоренных методов, получивший название *Каталист* (Catalyst). Немного подробнее об этом будет написано в 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 $\tilde{O}(n)$ (сложностью проектирования) и сложностью вычисления градиента $O(n^2)$.[16] 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 все равно, чтобы сделать шаг метода, нужно посчитать градиент, поэтому сложность «проектирования» вполне можно «утяжелять», например, за счет отмеченной идеи композитной оптимизации, до сложности расчета градиента. Общая сложность итерации по порядку сохранится, но зато число итераций может существенно уменьшиться. Продолжая движение в намеченном направлении, приведем другой пример задачи «со структурой», которая также позволяет заносить часть

[16] Такой большой зазор ($n \leftrightarrow n^2$) имеет место не всегда. Однако в типичных ситуациях вычисление градиента занимает значительно больше времени, чем последующее проектирование.

сложности задачи в «проектирование», сохраняя общую конструкцию [51], [52, § 4, гл. 7], [54, п. 4.3], [199, п. 4], [222]:

$$f(x) = F\left(f_1(x), ..., f_m(x)\right) \to \min\limits_{x \in Q}, \tag{3.23}$$

где все функции выпуклые, причем функция $F(y)$ еще и неубывающая по каждому из своих аргументов. Также предполагаем, что все функции $f_j(x)$, $j = 1, ..., m$ удовлетворяют условию (2.3) с $L = L_j$, $j = 1, ..., m$, а функция $F(y)$ удовлетворяет условиям (2.3) – (2.5) с $\nu = 0$ ($L = L_0$) и $\| \ \| = \| \ \|_1$. Подобно (3.2), (3.20), положим

$$\psi_\delta(y, x) = F\left(f_1(x) + \left\langle \nabla f_1(x), y - x \right\rangle, ..., f_m(x) + \left\langle \nabla f_m(x), y - x \right\rangle\right) - f(x). \tag{3.24}$$

Сделанные предположения позволяют утверждать, что условие (3.1) выполняется при $L = L_0 \sum\limits_{j=1}^{m} L_j$. Получаемый при таком выборе $\psi_\delta(y, x)$ (см. (3.24)) метод (3.3) называют *методом уровней* (level method).

Достаточно популярным частным случаем задачи (3.23) является задача, в которой $F(y) = \max\limits_{j=1,...,m} y_j$ [54, п. 2.3]. К такой задаче с помощью *метода нагруженного функционала* [11, § 19, гл. 5], [61, § 3, гл. 9] сводятся и задачи *условной оптимизации* (задачи с функциональными ограничениями) вида

$$f_0(x) \to \min\limits_{\substack{f_1(x) \le 0, ..., f_m(x) \le 0, \\ x \in Q}}. \tag{3.25}$$

Действительно, задачу (3.25) можно переписать следующим образом: найти такой $t = t_*$ и соответствующий $x(t_*)$, доставляющий решение вспомогательной задачи минимизации в (3.26), что $G(t) > 0$ при $t < t_*$ и $G(t_*) = 0$, где

$$G(t) = \min\limits_{x \in Q} \max\left\{f_0(x) - t, f_1(x), ..., f_m(x)\right\}. \tag{3.26}$$

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 [116, гл. 3]:

1) пусть $\tilde{F}(x,y)$ – выпуклая функция, как функция x, тогда функция

$$f(x)=\max_{y\in Q}\tilde{F}(x,y)$$

также выпуклая. Хорошей иллюстрацией тут является представление $|x|=\max\{-x,x\}$, $x\in\mathbb{R}$.

2) пусть $\bar{F}(x,y)$ – выпуклая функция, как функция (x,y), а $\bar{Q}$ – выпуклое множество, тогда функция

$$f(x)=\min_{y:\,(x,y)\in\bar{Q}}\bar{F}(x,y)$$

также выпуклая. Это следует из того, что пересечение надграфика выпуклой функции с выпуклым цилиндром с основанием Q также будет выпуклым множеством и его проекция вдоль y также будет выпуклым множеством. ■

Из общих результатов о поиске корня скалярного нелинейного уравнения [118, гл. 4] можно пытаться найти t_* с относительной точностью ε за $\mathrm{O}\left(\ln\left(\varepsilon^{-1}\right)\right)$ 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 приближенно, то и посчитать[17] $x(t)$ можно только приближенно. Это обстоятельство приводит к необходимости более тонкого анализа. Детали см., например, в [54, п. 2.3]. Однако сохраняется общий вывод о возрастании сложности решения задачи (3.25) по сравнению с (3.23) в $\mathrm{O}\left(\ln\left(\varepsilon^{-1}\right)\right)$ раз при рассматриваемом подходе. ■

Все последующие рассуждения могут проводиться в общности, выбранной в данном параграфе. Однако в методических целях далее мы намеренно не будем «гнаться за общностью» и 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-приближенное по функции решение $x_\varepsilon\in Q$, т. е.

---

[17] Здесь $x(t)$ – решение задачи вспомогательной задачи минимизации по $x\in Q$ в (3.26).

$$f(x_\varepsilon) - f(x_*) \le \varepsilon .$$

1) Пусть функция $f(x)$ удовлетворяет условию (2.27) при $\nu = 1$ с константой L_1. Покажите, что тогда имеет место следующая оценка:

$$\langle \nabla f(x_\varepsilon), x_\varepsilon - x_* \rangle \le \|x_\varepsilon - x_*\| \sqrt{2L_1\varepsilon} .$$

Пусть $R = \max\limits_{x,y\in Q} \|y - x\| < \infty$. Покажите, что тогда имеет место следующая оценка:

$$\max_{x\in Q} \langle \nabla f(x_\varepsilon), x_\varepsilon - x \rangle \le R\sqrt{2L_1\varepsilon} .$$

2) Пусть функция $f(x)$ удовлетворяет условию (2.27) при $\nu = 1$ с константой L_1, является μ-сильно выпуклой в норме $\|\ \|$ и $\nabla f(x_*) = 0$. Покажите, что тогда имеет место следующая оценка:

$$\langle \nabla f(x_\varepsilon), x_\varepsilon - x_* \rangle \le 2\varepsilon\sqrt{L_1/\mu} .$$

3) Пусть функция $f(x)$ удовлетворяет условию (2.27) при $\nu = 0$ с константой L_0 и является μ-сильно выпуклой в норме $\|\ \|$. Покажите, что тогда имеет место следующая оценка:

$$\langle \nabla f(x_\varepsilon), x_\varepsilon - x_* \rangle \le L_0\sqrt{2\varepsilon/\mu} .$$

Упражнение 3.2. Пусть $f(x) = \min\limits_{y\in\tilde{Q}} \bar{F}(y,x)$, где $\tilde{Q}$ – ограниченное выпуклое множество, а $\bar{F}(y,x)$ – такая достаточно гладкая, выпуклая по совокупности переменных функция, что при $y, y' \in \tilde{Q}$, $x, x' \in \mathbb{R}^n$:

$$\left\|\nabla\bar{F}(y',x') - \nabla\bar{F}(y,x)\right\|_2 \le L\left\|(y',x') - (y,x)\right\|_2 .$$

Пусть для произвольного x можно найти такой $\tilde{y}_\delta(x) \in \tilde{Q}$, что (следует сравнить с (3.4)):

$$\max_{y\in\tilde{Q}} \langle \nabla_y \bar{F}(\tilde{y}_\delta(x), x), \tilde{y}_\delta(x) - y \rangle \le \delta .$$

Покажите, что для любых $x, x' \in \mathbb{R}^n$

$$\bar{F}(\tilde{y}_\delta(x), x) - f(x) \le \delta,\ \left\|\nabla f(x') - \nabla f(x)\right\|_2 \le L\|x' - x\|_2$$

и

$$\left(\bar{F}(\tilde{y}_\delta(x), x) - 2\delta; \langle \nabla_y \bar{F}(\tilde{y}_\delta(x), x), y - x \rangle\right)$$

будет $(6\delta, 2L)$-моделью для функции $f(x)$ в точке x относительно 2-нормы.

Указание. См. [18]. Интересно сопоставить это упражнение с леммой 13 из п. 5 § 1, гл. 5 [61]. ■

Упражнение 3.3 (прокс-метод с неточным решением задачи минимизации на итерации). Рассмотрим функцию (следует сравнить с задачей из (3.21)):

$$f(x) = \min_{y \in \hat{Q}} \underbrace{\left\{ \varphi(y) + \frac{L}{2}\|y - x\|_2^2 \right\}}_{\Psi(y,x)}.$$

Предположим, что $\varphi(y)$ – выпуклая функция и

$$\max_{y \in \hat{Q}} \left\{ \Psi(y(x), x) - \Psi(y, x) + \frac{L}{2}\|y - y(x)\|_2^2 \right\} \le \delta .$$

Покажите, что тогда

$$\left(\varphi(y(x)) + \frac{L}{2}\|y(x) - x\|_2^2 - \delta; \left\langle L \cdot (x - y(x)), y - x \right\rangle \right)$$

будет (δ, L)-моделью функции $f(x)$ в точке x относительно 2-нормы.

Указание. См. [140]. Отметим также, что если

$$x_* \in \operatorname{Arg}\min_x f(x) \left(= \operatorname{Arg}\min_{x \in \hat{Q}} f(x) \right),$$

то тогда [61, теорема 5 п. 2 § 1, гл. 6]:

$$x_* \in \operatorname{Arg}\min_{y \in \hat{Q}} \varphi(y) . \blacksquare$$

Упражнение 3.4 (градиентное отображение). Изложенный в этом параграфе подход является далеко не единственным способом получения части описанных в § 3 результатов. Удобным инструментом также является использование *градиентного отображении*, см., например, [53], [54, п. 2.3]. С помощью градиентного отображения обобщается (путем замены градиента на градиентное отображение) основной набор базовых формул, из которых выводятся все последующие оценки,[18] см., например, [54,

[18] 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 определенному (явно выписываемому и конечному!) набору условий, связывающих значения функции и её (суб-)градиента в генерируемых методом (1.33) точках [275]. Это наблюдение позволяет получать

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. [21, 150, 215, 247]. ■

Упражнение 3.6 (Двуреченский–Нестеров, 2018). В работе [236] в связи с изучением процессов, происходящих в ходе избирательных компаний, и в связи с изучение быстрых способов кластеризации многомерных данных предлагается искать решение следующей задачи выпуклой оптимизации:

$$f_\mu\left(x=(z,p)\right)=g\underbrace{(z,p)}_{x}+\mu\sum_{k=1}^{n}z_k\ln z_k+\frac{\mu}{2}\|p\|_2^2\to\min_{z\in S_n(1),\ p\in\mathbb{R}_+^m}.$$

Введем норму $\|x\|^2=\|(z,p)\|^2=\|z\|_1^2+\|p\|_2^2$. Убедитесь, что $\|\ \|$ действительно норма. Предположим, что

$$\left\|\nabla g(x_2)-\nabla g(x_1)\right\|_*\le L\|x_2-x_1\|,$$

где $L\le\mu$. Покажите, что

$$\left(f_\mu(x);\ f_{\mu-L}(y)+\left\langle\nabla g(x),y-x\right\rangle-g(y)\right)$$

будет $(0,2L)$-моделью функции $f_\mu(x)$ в точке x относительно нормы $\|\ \|$. Заметим, что выпуклость или простота функции $g(x)$ здесь не требуется!

Указание. Идея такой модели была заимствована из работ [89, 92]. ■

Упражнение 3.7 (*метод подобных треугольников* [24, 72]). Для задачи (2.1) рассмотрите следующий вариант быстрого (ускоренного) градиентного спуска с одной проекцией, работающий с моделью функции (3.1):

$$x^0=u^0,\ A_0=\alpha_0=0,$$

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

$$A_{k+1}=A_k+\alpha_{k+1},$$

точные минимаксные оценки скорости сходимости различных итерационных процедур вида (1.33) [137, 275, 276], см. также замечание 1.5.

$$y^{k+1} = \frac{\alpha_{k+1} u^k + A_k x^k}{A_{k+1}},$$

$$u^{k+1} = \arg_{\tilde{\delta}} \min_{x \in Q} \left\{ \alpha_{k+1} \psi_\delta \left(x, y^{k+1} \right) + V\left(x, u^k \right) \right\},$$

$$x^{k+1} = \frac{\alpha_{k+1} u^{k+1} + A_k x^k}{A_{k+1}}.$$

Покажите, что

$$f\left(x^N\right) - f\left(x_*\right) = \mathrm{O}\left(\frac{LR^2}{N^2} + \frac{\tilde{\delta}}{N} + N\delta \right).$$

Полезно сравнить эту формулу с (3.15).

Покажите, что оценка скорости сходимости описанного метода не ухудшится, если на каждой итерации делать в конце дополнительное присваивание: в качестве x^{k+1} выбирать ту точку среди $\left\{ y^{k+1}, u^{k+1}, x^{k+1} \right\}$, которая доставляет наименьшее значение целевой (минимизиремой) 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 метод последовательности точек $\left\{ x^k \right\}_k$ целевая функция будет монотонно убывать.

§ 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) [226] (по А.С. Немировскому) или, в общем случае, в терминах неравенств типа (4.2) [15, 53, 54, 102, 235] (по Ю.Е. Нестерову).

В данном параграфе будет продемонстрирована прямодвойственная природа градиентного спуска [2]. Ниже мы постараемся пояснить, как, решая двойственную задачу методом (2.6) с шагом (2.21), можно с такой же точностью восстанавливать решение прямой задачи. Сначала планируется рассматривать двойственные задачи. Для двойственных задач множество Q – либо все 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).

◊ В действительности, при правильном взгляде [226], 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) [226]. 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 [226, 234]. ◊

Итак, вернемся к формуле (2.12) с $h = 1/L$ (2.21). Перепишем её следующим образом:

$$f\left(x^{k+1}\right) \le \left\{ f\left(x^k\right) + \left\langle \nabla f\left(x^k\right), x - x^k \right\rangle \right\} + \frac{L}{2}\left\|x - x^k\right\|_2^2 - \frac{L}{2}\left\|x - x^{k+1}\right\|_2^2 + \delta \ . \quad (4.1)$$

Суммируя (4.1) по $k=0,...,N-1$, учитывая выпуклость функции $f(x)$ и произвол в выборе $x\in Q$, получим

$$f\left(\overline{x}^N\right)\le\frac{1}{N}\min_{x\in Q}\left\{\sum_{k=0}^{N-1}\left[f\left(x^k\right)+\left\langle\nabla f\left(x^k\right),x-x^k\right\rangle\right]+\frac{L}{2}\left\|x-x^0\right\|_2^2\right\}+\delta, \quad (4.2)$$

где

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

Данная формула является обоснованием *прямодвойственности* метода градиентного спуска (2.6), (2.21) [15, гл. 3], [2, 226, 234]. 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 сопряженной (двойственной) задачи. Далее в этом параграфе на примере задачи минимизации выпуклой функции при аффинных ограничениях демонстрируется сначала как прямодвойственный метод применяется к двойственной задаче (решение прямой задачи удается восстановить за счет прямодвойственности метода), а затем (в конце параграфа) градиентный спуск будет применен к исходной (прямой) задаче при дополнительном предположении, что аффинные ограничения имеют достаточно простую структуру (решение двойственной задачи также удается восстановить за счет прямодвойственности метода). Отмеченные возможности прямодвойственных методов, рассмотренные далее на примере только градиентного спуска, отчасти и объясняют их название [234].

Рассмотрим следующий (вычислимый! – ввиду простоты множества Q, см. § 2) критерий останова метода:

$$f\left(\overline{x}^N\right)-\frac{1}{N}\min_{x\in B_{R,Q}\left(x^0\right)}\left\{\sum_{k=0}^{N-1}\left[f\left(x^k\right)+\left\langle\nabla f\left(x^k\right),x-x^k\right\rangle\right]\right\}\le\varepsilon, \quad (4.3)$$

который получается из (4.2), (2.18) при $\delta=\varepsilon/2$. Обратим внимание, что минимум в (4.3) берется по множеству $B_{R,Q}\left(x^0\right)$, где $R=\left\|x_*-x^0\right\|_2$, а не $B_{R,Q}\left(x_*\right)$, поскольку x_* нам не известно.[19] При этом ранее в § 2 мы пока-

[19] Строго говоря, и $R=\left\|x_*-x^0\right\|_2$ также неизвестен. Однако при использовании $B_{R,Q}\left(x^0\right)$ удается ограничиться лишь одним неизвестным R, а в ряде случаев достаточно здесь исходить из размера множества Q.

зали, что $x_* \in B_{R,Q}\left(x^0\right)$. Значит по выпуклости $f\left(x\right)$ (см. (2.13)) имеем нужное нам неравенство

$$\begin{aligned} f\left(\overline{x}^N\right) - f\left(x_*\right) \le f\left(\overline{x}^N\right) - \frac{1}{N}\sum_{k=0}^{N-1}\left[f\left(x^k\right) + \left\langle \nabla f\left(x^k\right), x_* - x^k \right\rangle \right] \le \\ \le f\left(\overline{x}^N\right) - \frac{1}{N}\min_{x \in B_{R,Q}\left(x^0\right)} \left\{ \sum_{k=0}^{N-1}\left[f\left(x^k\right) + \left\langle \nabla f\left(x^k\right), x - x^k \right\rangle \right] \right\} \le \varepsilon. \end{aligned} \tag{4.4}$$

С другой стороны, из (4.2) имеем

$$\begin{aligned} f\left(\overline{x}^N\right) \le \frac{1}{N}\min_{x \in Q} \left\{ \sum_{k=0}^{N-1}\left[f\left(x^k\right) + \left\langle \nabla f\left(x^k\right), x - x^k \right\rangle \right] + \frac{L}{2}\left\| x - x^0 \right\|_2^2 \right\} + \frac{\varepsilon}{2} \le \\ \le \frac{1}{N}\min_{x \in B_{R,Q}\left(x^0\right)} \left\{ \sum_{k=0}^{N-1}\left[f\left(x^k\right) + \left\langle \nabla f\left(x^k\right), x - x^k \right\rangle \right] \right\} + \frac{LR^2}{2N} + \frac{\varepsilon}{2}. \end{aligned} \tag{4.5}$$

Значит, с учетом (2.5), (2.19) метод (2.6), (2.21) при условии (2.4) гарантированно остановится по критерию (4.3), сделав не более

$$N = \frac{LR^2}{\varepsilon} \le \left(\frac{L_\nu R^{1+\nu}}{\varepsilon} \right)^{\frac{2}{1+\nu}} \tag{4.6}$$

итераций (вычислений $\nabla f\left(x^k\right)$).

Рассмотрим конкретный пример использования оценки типа (4.2) [2, 139, 154, 277]. 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*):

$$\varphi\left(y\right) \to \min_{Ay=b,\ y \in \tilde{Q}}, \tag{4.7}$$

где функция $\varphi\left(y\right)$ – μ-сильно выпуклая в *p*-норме на $\tilde{Q}$ $\left(1 \le p \le 2\right)$. Решение задачи (4.7) обозначим через y_*, а оптимальное значение функционала – через φ_* ($\varphi_* = \varphi\left(y_*\right)$).

Построим (с точностью до знака) *двойственную задачу* к задаче (4.7):

$$f\left(x\right) = \max_{y \in \tilde{Q}} \left\{ \left\langle x, b - Ay \right\rangle - \varphi\left(y\right) \right\} \to \min_{x \in \mathbb{R}^n}. \tag{4.8}$$

◊ Опишем общий принцип построения двойственных задач [116, гл. 5]. Итак, пусть исходная задача выпуклой оптимизации имеет вид

$$\varphi\left(y\right) \to \min_{h(y) \le 0,\ Ay=b,\ y \in \tilde{Q}}.$$

Тогда

$$\min_{h(y)\le 0,\ Ay=b,\ y\in\tilde{Q}} \varphi(y) = \min_{y\in\tilde{Q}} \left\{ \varphi(y) + \max_{z\ge 0} \langle z, h(y)\rangle + \max_{x} \langle x, Ay-b\rangle \right\} \stackrel{?}{=}$$

$$\stackrel{?}{=} \max_{z\ge 0,\ x} \min_{y\in\tilde{Q}} \left\{ \varphi(y) + \langle z, h(y)\rangle + \langle x, Ay-b\rangle \right\}.$$

Равенство со знаком вопроса обосновывается с помощью теорем типа *фон Неймана* или *Сиона–Какутани* [223, приложение D.4]. К сожалению, при таком подходе требуется компактность множества Q или возможность компактифицировать двойственные переменные (z, x) (см. замечание 4.2 и упражнение 4.1). В любом случае в реальных задачах, как правило, удается обосновать это равенство [116], которое также называют *сильной двойственностью* [116, гл. 5]. Таким образом, решение исходной задачи сводится к двойственной задаче (с точностью до знака):

$$\max_{y\in\tilde{Q}} \left\{ \langle x, b-Ay\rangle - \langle z, h(y)\rangle - \varphi(y) \right\} \to \min_{z\ge 0,\ x}. \ \Diamond$$

Точное решение задачи (4.8) будем обозначать через $y(x)$. Во многих важных приложениях основной вклад в сложность расчета $y(x)$ дает умножение Ay. Это так, например, для сепарабельных функционалов

$$\varphi(y) = \sum_{i=1}^{m} \varphi_i(y_i)$$

и параллелепипедных ограничений $\tilde{Q}$. 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 (4.8), выполняется условие (2.4) с $\nu = 1$ и $L_1 = L = \frac{1}{\mu} \max_{\|y\|_p \le 1} \|Ay\|_2^2$ [2, 235]. В частности, для $p = 1$

$$L = \frac{1}{\mu} \max_{j=1,\dots,m} \left\| A^j \right\|_2^2,$$

где A^j – j-й столбец матрицы A. Для $p = 2$

$$L = \frac{1}{\mu}\lambda_{\max}\left(A^T A\right) \overset{def}{=} \frac{1}{\mu}\sigma_{\max}\left(A\right).$$

Замечание 4.1 (метод регуляризации и техника двойственного сглаживания). Добиться сильной выпуклости $\varphi(y)$ всегда можно с помощью *регуляризации* задачи. Опишем, в чем состоит *техника регуляризации* (см., например, [4], [12, гл. 9] и цитированную там литературу), восходящая к работам трех отечественных научных школ: А.Н. Тихонова (Москва), М.М. Лаврентьева (Новосибирск), В.К. Иванова (Свердловск), занимавшихся изучением некорректных задач. Рассмотрим новую задачу:

$$\varphi^{\mu}\left(y\right) = \varphi\left(y\right) + \mu V\left(y, y^0\right) \to \min_{Ay=b,\ y\in\tilde{Q}}, \tag{4.9}$$

где $V\left(y, y^0\right)$ – 1-сильно выпуклая в p-норме функция y. Обозначим через φ_*^{μ} оптимальное значение функционала в задаче (4.9). Пусть[20]

$$\mu \le \frac{\varepsilon}{2V\left(y_*, y^0\right)}, \tag{4.10}$$

и удалось найти $\varepsilon/2$-решение задачи (4.9), т. е. нашелся такой $y_{\varepsilon/2}$, что $Ay_{\varepsilon/2} = b$, $y_{\varepsilon/2} \in \tilde{Q}$:

$$\varphi^{\mu}\left(y_{\varepsilon/2}\right) - \varphi_*^{\mu} \le \varepsilon/2.$$

Тогда

$$\varphi\left(y_{\varepsilon/2}\right) - \varphi_* \le \varepsilon.$$

Действительно,

$$\varphi\left(y_{\varepsilon/2}\right) - \varphi_* \le \varphi^{\mu}\left(y_{\varepsilon/2}\right) - \varphi_* \le \varphi^{\mu}\left(y_{\varepsilon/2}\right) - \varphi_*^{\mu} + \varepsilon/2 \le \varepsilon.$$

Здесь использовались определение φ_*^{μ} и формула (4.10):

$$\varphi_*^{\mu} = \min_{Ay=b,\ y\in\tilde{Q}}\left\{\varphi\left(y\right) + \mu V\left(y, y^0\right)\right\} \le \varphi\left(y_*\right) + \mu V\left(y_*, y^0\right) \le \varphi_* + \varepsilon/2.$$

---

[20] Как правило, величина $V\left(y_*, y^0\right)$ неизвестна, поэтому на практике используются рестарты по параметру μ, приводящие к увеличению общего числа итераций в несколько раз [16, 88], см. также указание к упражнению 2.3.

Стоит отметить, что если изначально рассматривалась задача вида (4.8), то говорят, что функционал $f(x)$ представим в *форме Лежандра*. Пусть $\tilde{Q}$ – 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 $\tilde{R}^2 = \max\limits_{y \in \tilde{Q}} V(y, y^0)$ с $\mu \le \varepsilon / (2\tilde{R}^2)$, приводит к сглаживанию функции

$$f(x) \to f_{\mu}(x) = \max_{y \in \tilde{Q}} \{\langle x, b - Ay \rangle - \varphi(y) - \mu V(y, y^0)\}, \quad 0 \le f(x) - f_{\mu}(x) \le \varepsilon/2. \tag{4.11}$$

При этом $f_{\mu}(x)$ будет иметь константу Липшица градиента в 2-норме:

$$L_{\varepsilon} = \frac{2\tilde{R}^2}{\varepsilon} \max_{\|y\|_p \le 1} \|Ay\|_2^2.$$

Простейший пример такого сглаживания:

$$f(x) = \max_{l=1,\ldots,m} \langle c^l, x \rangle = \max_{y \in S_m(1)} \sum_{l=1}^{m} y_l \langle c^l, x \rangle \to$$

$$\to \max_{y \in S_m(1)} \left\{ \sum_{l=1}^{m} y_l \langle c^l, x \rangle - \mu \sum_{l=1}^{m} y_l \ln \left(\frac{y_l}{1/m} \right) \right\} =$$

$$= \mu \ln \left(\sum_{l=1}^{m} \exp \left(\langle c^l, x \rangle / \mu \right) \right) - \mu \ln m = f_{\mu}(x),$$

где $\mu = \varepsilon / (2 \ln m)$. Описанную выше конструкцию (4.11) обычно называют *двойственным сглаживанием* или *техникой сглаживания по Нестерову* [53, гл. 5], [235]. В классе рассматриваемых в этой главе неускоренных методов данная техника по оценкам не дает 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 [53, гл. 5], [235]:

$$\sim \sqrt{L_{\varepsilon} \varepsilon^{-1}} \sim \sqrt{\varepsilon^{-2}} \sim \varepsilon^{-1}.$$

Разумеется, имеет смысл говорить о двойственном сглаживании только в случае, когда задача максимизации в (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 определенных (композитных) случаях описанная техника позволяет получать новые результаты, недостижимые с помощью техники замечания 5.1, см., например, [130, 147]. Отметим также, что есть и другие способы сглаживания (см., например, [88]), впрочем, также имеющие весьма ограниченную область применимости.

Хорошо известный пример использования регуляризации – способ вычисления (понимания) *псевдообратной матрицы* [61, 117]:

$$A^+ = \lim_{\mu \to 0+} \left(A^T A + \mu I \right)^{-1} A^T .$$

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 [76, лекции 13, 14]. На рассмотренном примере также удобно демонстрировать связь метода регуляризации и метода *штрафных функций*, см. замечание 4.3 и [97].

Менее известный, но не менее интересный пример *итеративной регуляризации/сглаживания* имеется в работе [165], в которой решение седловой билинейной задачи сводится к последовательности задач выпуклой оптимизации в условиях острого минимума [61, 259]. ■

◊ Так же как и в упражнении 2.3, здесь можно 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. ◊

Положим[21] $x^0 = 0$, $N = 2LR^2/\varepsilon$, где $R = \|x_*\|_2$. Подобно (4.5) можно написать:

[21] Как ни странно, такой выбор точки старта является существенным в последующих рассуждениях. Нам не известен способ рассуждений, который бы

$$f\left(\bar{x}^N\right) \le \frac{1}{N}\min_{x\in\mathbb{R}^n}\left\{\sum_{k=0}^{N-1}\left[f\left(x^k\right)+\left\langle\nabla f\left(x^k\right),x-x^k\right\rangle\right]+\frac{L}{2}\left\|x-x^0\right\|_2^2\right\}\le$$

$$\le \frac{1}{N}\min_{x\in B_{2R}(0)}\left\{\sum_{k=0}^{N-1}\left[f\left(x^k\right)+\left\langle\nabla f\left(x^k\right),x-x^k\right\rangle\right]\right\}+\underbrace{\frac{2LR^2}{N}}_{\varepsilon}.$$

Ввиду (4.8) отсюда *по формуле Демьянова–Данскина* [32, 33, 110]

$$\nabla f\left(x\right)=b-Ay\left(x\right)$$

имеем

$$f\left(\bar{x}^N\right)-\frac{1}{N}\sum_{k=0}^{N-1}\left\langle x^k,b-Ay\left(x^k\right)\right\rangle+\frac{1}{N}\sum_{k=0}^{N-1}\varphi\left(y\left(x^k\right)\right)-$$

$$-\frac{1}{N}\min_{x\in B_{2R}(0)}\left\{\sum_{k=0}^{N-1}\left\langle b-Ay\left(x^k\right),x-x^k\right\rangle\right\}\le\varepsilon. \qquad (4.12)$$

По выпуклости $\varphi\left(y\right)$ из (4.12) имеем

$$f\left(\bar{x}^N\right)+\varphi\underbrace{\left(\frac{1}{N}\sum_{k=0}^{N-1}y\left(x^k\right)\right)}_{\bar{y}^N}+\max_{x\in B_{2R}(0)}\left\{\left\langle A\frac{1}{N}\sum_{k=0}^{N-1}y\left(x^k\right)-b,x\right\rangle\right\}\le\varepsilon,$$

т. е.

$$f\left(\bar{x}^N\right)+\varphi\left(\bar{y}^N\right)+2R\left\|A\bar{y}^N-b\right\|_2\le\varepsilon. \qquad (4.13)$$

Из (4.13) и *слабой двойственности* $-\varphi\left(y_*\right)\le f\left(x_*\right)$ имеем

$$\varphi\left(\bar{y}^N\right)-\varphi\left(y_*\right)\le\varphi\left(\bar{y}^N\right)+f\left(\bar{x}^N\right)\le f\left(\bar{x}^N\right)+\varphi\left(\bar{y}^N\right)+2R\left\|A\bar{y}^N-b\right\|_2\le\varepsilon.$$

◊ Фактически слабая двойственность – это отражение простого факта, что всегда имеет место неравенство

$$\max_y\min_x L\left(x,y\right)\le\min_x\max_y L\left(x,y\right).$$

На самом деле во всех естественных ситуациях, когда рассматриваются невырожденные (совместные) выпуклые задачи, в этом неравенстве имеет место равенство, т. е. имеет место сильная двойственность [116, гл. 5]. ◊

Также из (4.8), (4.13) можно получить, что

позволял получить аналогичный результат (теорема 4.1) без данного предположения: $x^0=0$.

$$R\left\|A\bar{y}^N-b\right\|_2 \le \overbrace{\underbrace{\left\langle \bar{x}^N, b-A\bar{y}^N\right\rangle}_{\ge -R\left\|A\bar{y}^N-b\right\|_2}-\varphi\left(\bar{y}^N\right)}^{\le f\left(\bar{x}^N\right)}+\varphi\left(\bar{y}^N\right)+2R\left\|A\bar{y}^N-b\right\|_2 \le \varepsilon .$$

Таким образом, установлен следующий результат.

Теорема 4.1. *Пусть нужно решить задачу* (4.7) *в следующем смысле*

$$\varphi\left(\bar{y}^N\right)-\varphi\left(y_*\right)\le\varepsilon,\ \left\|A\bar{y}^N-b\right\|_2\le\tilde{\varepsilon}. \tag{4.14}$$

Для этого рассмотрим двойственную задачу (4.8), *которую будем решать градиентным спуском* (1.22):

$$x^{k+1}=x^k-\frac{1}{L}\nabla f\left(x^k\right)$$

с $x^0=0$. *Выберем в качестве критерия останова метода следующие условия* (*зазор двойственности и невязку в ограничении*):

$$f\left(\bar{x}^N\right)+\varphi\left(\bar{y}^N\right)\le\varepsilon,\ \left\|A\bar{y}^N-b\right\|_2\le\tilde{\varepsilon},$$

где

$$\bar{x}^N=\frac{1}{N}\sum_{k=1}^{N}x^k,\ \bar{y}^N=\frac{1}{N}\sum_{k=0}^{N-1}y\left(x^k\right),$$

из которых вытекает (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 $L=\frac{1}{\mu}\max\limits_{\|y\|_p\le 1}\|Ay\|_2^2$, $R=\|x_*\|_2$. *Если решение задачи* (4.8) x_* *не единственно, то в оценке* R *в* (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 $R=\|x_*\|_2$ (4.8). Если решение x_* не единственно, то выбирается наименьшее по 2-норме (см. § 1, 2). Это R можно оценить следующим образом [53, п. 4.3.4], [201]:

$$R^2=\|x_*\|_2^2\le\left\|\nabla\varphi\left(y_*\right)\right\|_2^2\Big/\tilde{\sigma}_{\min}\left(A\right), \tag{4.16}$$

где

$$\tilde{\sigma}_{\min}(A) = \min\left\{\lambda > 0 : \exists\ x \neq 0 : AA^T x = \lambda x\right\}.$$

Действительно, исходя из определения x_* и y_*, имеем для любого $y \in \tilde{Q}$

$$-\varphi(y_*) = \langle x_*, b - Ay_* \rangle - \varphi(y_*) = \max_{y \in \tilde{Q}} \left\{\langle x_*, b - Ay \rangle - \varphi(y)\right\} \geq \langle x_*, b - Ay \rangle - \varphi(y),$$

т. е.

$$-\varphi(y_*) = \langle x_*, b - Ay_* \rangle - \varphi(y_*) = \max_{y \in \tilde{Q}} \left\{\langle x_*, b - Ay \rangle - \varphi(y)\right\} \geq$$

$$\geq \langle x_*, b - Ay \rangle - \varphi(y) = \langle x_*, Ay_* - Ay \rangle - \varphi(y) = \langle A^T x_*, y_* - y \rangle - \varphi(y).$$

Следовательно, для любого $y \in \tilde{Q}$

$$\varphi(y) \geq \varphi(y_*) + \langle -A^T x_*, y - y_* \rangle.$$

По выпуклости $\varphi(y)$ отсюда следует, что

$$-A^T x_* = \nabla \varphi(y_*).$$

Точнее, $-A^T x_* \in \partial \varphi(y_*)$. Осталось только заметить, для любого $x \in \left(\operatorname{Ker} A^T\right)^{\perp}$ имеет место неравенство

$$\left\|-A^T x\right\|_2^2 = \langle -A^T x, -A^T x \rangle = \langle x, AA^T x \rangle \geq \tilde{\sigma}_{\min}(A) \|x\|_2^2. \blacksquare$$

Пример 4.1 (децентрализованная распределенная оптимизация [201, 221, 261, 281]). 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 $\varphi_i''(y) \geq \mu$, $i = 1, \ldots, n$, $y \in \mathbb{R}$. Предположим, что есть связанная сеть $G = \langle V, E \rangle$ из n узлов. В i-м узле хранится функция $\varphi_i(y)$. Зададим матрицу инциденций $\mathrm{I} = \left\|\mathrm{I}_{ij}\right\|_{i,j=1}^{n}$: $\mathrm{I}_{ij} = 1$, $(i, j) \in E$; $\mathrm{I}_{ij} = 0$, $(i, j) \neq E$. По матрице I построим симметричную неотрицательно определенную *лапласову матрицу* $W \succ 0$:

$$W_{ij} = \begin{cases} -\mathrm{I}_{ij}, i \neq j, \\ \sum\limits_{j=1}^{n} \mathrm{I}_{ij}, i = j. \end{cases}$$

По *теореме Фробениуса–Перрона* [56, § 7, 8, гл. 2]:

$$Wy = 0 \;\Leftrightarrow\; y_1 = \ldots = y_n . \tag{4.18}$$

Ввиду (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. (4.8)):

$$\begin{aligned} f(x) = \varphi^*(-Wx) = \max_{y \in \mathbb{R}^n} \left\{ -\langle x, Wy \rangle - \varphi(y) \right\} = \\ = \sum_{i=1}^{n} \max_{y_i \in \mathbb{R}} \left\{ \left[-Wx \right]_i y_i - \varphi_i(y_i) \right\} \to \min_{x \in \mathbb{R}^n} . \end{aligned} \tag{4.20}$$

Считаем, что *i*-я вспомогательная задача максимизации в (4.20) может эффективно решаться в *i*-м узле. Заметим, что для функции $f(x)$ константу Липшица градиента можно оценить как $L = \sigma_{\max}(W)/\mu$ [2, 235], а на размер двойственного решения есть оценка (см. замечание 4.2) $R^2 = \|x_*\|_2^2 \le \|\nabla \varphi(y_*)\|_2^2 / \tilde{\sigma}_{\min}(W)$. Согласно написанному ранее в этом параграфе решать задачу (4.20) можно методом

$$x_i^{k+1} = x_i^k - \frac{1}{L} \left[\nabla f(x^k) \right]_i = x_i^k + \frac{1}{L} \left[W \tilde{y}(-Wx^k) \right]_i , \tag{4.21}$$

где через $\tilde{y}(-Wx)$ обозначается решение задачи (4.20). Итак, пусть в каждом узле хранятся $\left\{ x_i^k, \tilde{y}_i\left(\left[-Wx^k \right]_i \right) \right\}_k$. Ключевое наблюдение: чтобы вычислить $\tilde{y}_i\left(\left[-Wx^k \right]_i \right)$ *i*-у узлу необходимо обратиться только к своим непосредственным соседям за соответствующими компонентами вектора x^k (см. (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. (4.21)). Таким образом, один шаг градиентного спуска для двойственной задачи приводит к коммуникации каждого узла со своими соседями два раза

(передается два числа). Поскольку вычислительные возможности узлов, как правило, на несколько порядков выше скорости передачи информации по сети, то полученный дисбаланс (решать вспомогательную задачу поиска $\tilde{y}_i\left(\left[-Wx^k\right]_i\right)$ заметно труднее, чем послать и принять несколько чисел) хорошо способствует эффективному решению задачи.

Несложно заметить, что время работы алгоритма будет прямо пропорционально числу обусловленности матрицы W^2, т. е. $\sigma_{\max}(W)/\tilde{\sigma}_{\min}(W)$. В действительности, можно улучшить описанный выше подход, если сделать замену $W \to \sqrt{W}$ [261]:

$$\sqrt{W}y = 0 \;\Leftrightarrow\; y_1 = \ldots = y_n,$$

$$\tilde{y}\left(-\sqrt{W}x\right) = \arg\max_{y\in\mathbb{R}^n}\left\{-\left\langle\sqrt{W}x, y\right\rangle - \varphi(y)\right\},$$

$$\sqrt{W}x^{k+1} = \sqrt{W}x^k + \frac{1}{L}W\tilde{y}\left(-\sqrt{W}x^k\right).$$

Обозначая $z = \sqrt{W}x$, запишем метод в новых переменных:

$$\tilde{y}(z) = \arg\max_{y\in\mathbb{R}^n}\left\{\langle z, y\rangle - \varphi(y)\right\},$$

$$z^{k+1} = z^k + \frac{1}{L}W\tilde{y}\left(-z^k\right).$$

Легко понять, что такой метод также может работать распределено [201, 221, 261, 281]. Таким образом, можно редуцировать

$$\sigma_{\max}(W)/\tilde{\sigma}_{\min}(W) \text{ к } \sigma_{\max}\left(\sqrt{W}\right)/\tilde{\sigma}_{\min}\left(\sqrt{W}\right) = \sqrt{\sigma_{\max}(W)/\tilde{\sigma}_{\min}(W)}.$$

Последняя величина оценивается снизу диаметром графа $G = \langle V, E\rangle$ и во многих случаях имеет тот же порядок (см., например, [281] и цитированную там литературу).

На основе ускоренных градиентных методов можно построить более быстрые децентрализованные распределенные алгоритмы решения задачи (4.17), см. [261, 281].

К сожалению, во всех случаях (ускоренном и неускоренном) не удается построить адаптивные / универсальные (см. § 5) варианты таких методов. Равно как и не удается предложить эффективный (практический) критерий останова методов (см. замечание 2.1 и § 4).

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 [127] (см. также [115, п. 7.3.1]). ■

◊ После работы [115] распределенная оптимизация (Grid-технологии) прочно закрепилась в современном анализе данных, см., например, концепцию Google: Federative Learning [196, 206, 216].

Хорошей практической демонстрацией, описанной в примере 4.1 техники, является распределенный способ вычисления барицентра Вассерштейна, учитывающий наличие явного *представления Лежандра* (сопряженного представления) для расстояния Монжа–Канторовича–Вассерштейна [134, 154, 251, 280].

Подобно теореме 1 можно распространить градиентный спуск и на невыпуклые задачи распределенной оптимизации [273]. ◊

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 [234]. Немного обобщим постановку задачи (4.7):

$$\varphi(y) = F(y) + g(y) \to \min_{Ay \le b,\ y \in \hat{Q}}. \tag{4.22}$$

Вместо равенства $Ay = b$ в (4.7) стали рассматривать неравенство $Ay \le b$ в (4.22) и добавили простой выпуклый композитный член $g(y)$. Будем предполагать, что выпуклая функция $F(y)$ удовлетворяет (только) условию (2.3) (см. также (2.26)), которое в данном случае будет иметь вид

$$F(y) \le F(z) + \langle \nabla F(z), y - z \rangle + \frac{L}{2}\|y - z\|^2 + \delta.$$

Рассмотрим метод вида (3.19) с шагом $h = 1/L$ (2.21) для задачи (4.22):

$$y^{k+1} = \arg\min_{Ay \le b,\ y \in \hat{Q}} \left\{ \langle \nabla F(y^k), y - y^k \rangle + g(y) + LV(y, y^k) \right\}. \tag{4.23}$$

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 $Ay \le b$ имеет вид $y \le \bar{y}$ или $y \ge \bar{y}$ [15, гл. 1, 3].

Повторяя рассуждения примера 3.1 (см. формулы (3.12), (3.20)), из (4.23) получим подобно оценке (4.2):

$$\varphi\left(\overline{y}^N\right) \le$$

$$\le \min_{Ay\le b,\ y\in\tilde{Q}} \left\{ \frac{1}{N}\sum_{k=0}^{N-1}\left[F\left(y^k\right) + \left\langle \nabla F\left(y^k\right), y-y^k \right\rangle + g\left(y\right)\right] + \frac{LV\left(y,y^0\right)}{N} \right\} + \delta, \tag{4.24}$$

где

$$\overline{y}^N = \frac{1}{N}\sum_{k=1}^{N} y^k .$$

Обозначим множитель Лагранжа к ограничению $Ay \le b$ в (4.24) через $\tilde{x}^N \ge 0$. Тогда (4.24) можно переписать следующим образом: для любого $\tilde{y} \in \tilde{Q}$

$$\varphi\left(\overline{y}^N\right) \le \min_{y\in\tilde{Q}} \left\{ \frac{1}{N}\sum_{k=0}^{N-1}\underbrace{\left[F\left(y^k\right) + \left\langle \nabla F\left(y^k\right), y-y^k \right\rangle + g\left(y\right)\right]}_{\le \varphi(y)} + \right.$$

$$\left. + \left\langle \tilde{x}^N, Ay-b \right\rangle + \frac{LV\left(y,y^0\right)}{N} \right\} + \delta \le \varphi\left(\tilde{y}\right) + \left\langle \tilde{x}^N, A\tilde{y}-b \right\rangle + \frac{LV\left(\tilde{y},y^0\right)}{N} + \delta. \tag{4.25}$$

Введем, подобно (4.8), двойственную (с точностью до знака) функцию

$$f\left(x\right) = \max_{y\in\tilde{Q}} \left\{ \left\langle x, b-Ay \right\rangle - \varphi\left(y\right) \right\}. \tag{4.26}$$

Обозначим, как и раньше, через $y\left(x\right)$ решение задачи максимизации в (4.26). Тогда, выбирая в (4.25) $\tilde{y} = y\left(\tilde{x}^N\right)$ и обозначая через $R^2 = V\left(y\left(\tilde{x}^N\right), y^0\right)$, получим

$$0 \le \varphi\left(\overline{y}^N\right) - \varphi\left(y_*\right) \le \varphi\left(\overline{y}^N\right) + f\left(\tilde{x}^N\right) \le \frac{LR^2}{N} + \delta . \tag{4.27}$$

◊ В отличие от (4.14) в (4.27) используется допустимая точка $\overline{y}^N$: $A\overline{y}^N \le b$, поэтому в (4.27) имеет место оценка снизу $0 \le \varphi\left(\overline{y}^N\right) - \varphi\left(y_*\right)$. Из слабой двойственности имеем

$$\varphi\left(\overline{y}^N\right) - \varphi\left(y_*\right) + f\left(\overline{x}^N\right) - f\left(x_*\right) \le \varphi\left(\overline{y}^N\right) + f\left(\overline{x}^N\right).$$

С учетом этих неравенств из (4.27) имеем

$$0 \le f\left(\overline{x}^N\right) - f\left(x_*\right) \le \varphi\left(\overline{y}^N\right) + f\left(\overline{x}^N\right) \le \frac{LR^2}{N} + \delta \,. \lozenge$$

Из формулы (4.27) (с $\delta = \varepsilon/2$) при условии, что функция $F\left(y\right)$ удовлетворяет (только) условию (2.27), следует, что метод вида (3.19) гарантированно остановится по критерию

$$\varphi\left(\overline{y}^N\right) + f\left(\overline{x}^N\right) \le \varepsilon \,,$$

сделав не более

$$N = \frac{2LR^2}{\varepsilon} \le \left(\frac{2L_\nu R^{1+\nu}}{\varepsilon}\right)^{\frac{2}{1+\nu}} \tag{4.28}$$

итераций (вычислений $\nabla F\left(y^k\right)$).

В заключение подчеркнем, в чем 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).

Упражнение 4.1 (условие Слейтера). Рассматривается задача выпуклой оптимизации

$$f\left(x\right) \to \min_{h(x) \le 0,\ x \in Q} \,,$$

где $h\colon\ \mathbb{R}^n \to \mathbb{R}^m$. Двойственная задача (с точностью до знака) имеет вид

$$\varphi\left(y\right) = \max_{x \in Q} \left\{-\left\langle y, h\left(x\right)\right\rangle - f\left(x\right)\right\} \to \min_{y \in \mathbb{R}^m_+} . \tag{4.29}$$

Обозначим через y^* решение двойственной задачи. Предположим, что выполняется *условие Слейтера*:

существует такая точка $\overline{x} \in Q$, *что* $h\left(\overline{x}\right) < 0$.

Пусть $\gamma = \min\limits_{i=1,\dots,m} \left\{-h_i\left(\overline{x}\right)\right\}$. Покажите, что

$$\left\|y^*\right\|_1 \le \frac{1}{\gamma}\left(f\left(\overline{x}\right) + \varphi(0)\right) = \frac{1}{\gamma}\left(f\left(\overline{x}\right) - \min_{x \in Q} f\left(x\right)\right).$$

Указание. См. [111]. Ключевое неравенство:

$$\varphi(0) \ge \varphi\left(y^*\right) = \max_{x \in Q}\left\{-\sum_{i=1}^m y_i^* h_i\left(x\right) - f\left(x\right)\right\} \ge -\sum_{i=1}^m y_i^* h_i\left(\overline{x}\right) - f\left(\overline{x}\right).$$

Аналогичным образом можно получать оценки на размер решения двойственной задачи и в более общих случаях (см., например, [22]). ■

Упражнение 4.2. Пусть задача из упражнения 4.1 решена в следующем смысле: найден такой $\tilde{y} \in \mathbb{R}^m_+$, что

$$\left\langle \tilde{y}, \nabla\varphi(\tilde{y}) \right\rangle \le \varepsilon,\ \left\| \left[-\nabla\varphi(\tilde{y}) \right]_+ \right\|_2 \le \tilde{\varepsilon},$$

где по формуле Демьянова–Данскина [116, гл. 3] $\nabla\varphi(\tilde{y}) = -h\left(x(\tilde{y})\right)$, $x(\tilde{y})$ – решение вспомогательной задачи максимизации в (4.29). Тогда

$$f\left(x(\tilde{y})\right) - f(x_*) \le \varepsilon,\ \left\| \left[h\left(x(\tilde{y})\right) \right]_+ \right\|_2 \le \tilde{\varepsilon}.$$

Указание. Ключевая выкладка:

$$-\left\langle \tilde{y}, h\left(x(\tilde{y})\right) \right\rangle - f\left(x(\tilde{y})\right) \ge -\langle \underbrace{\tilde{y}}_{\tilde{y}\ge 0}, \underbrace{h(x_*)}_{h(x_*)\le 0} \rangle - f(x_*) \ge -f(x_*).$$

Заметим также, что если здесь вместо ограничения в виде неравенства $h(x) \le 0$ имели бы аффинное ограничение в виде равенства $Ax - b = 0$, то тогда в проведенных рассуждения следовало бы сделать следующую корректировку: $\nabla\varphi(\tilde{y}) = b - Ax(\tilde{y})$, как следствие, условие $\left\|\nabla\varphi(\tilde{y})\right\|_2 \le \tilde{\varepsilon}$ обеспечивает выполнение условия $\left\|Ax(\tilde{y}) - b\right\|_2 \le \tilde{\varepsilon}$. ■

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 $\varphi(y)$ – μ-сильно выпуклая относительно p-нормы ($1 \le p \le 2$). Покажите, что функция $f(x)$ будет гладкой, с константой Липшица градиента в 2-норме:

$$L = \frac{1}{\mu} \max_{\|z\|_p \le 1} \|Az\|_2^2.$$

Более того, если $y_\delta(x)$ – решение вспомогательной задачи максимизации с точностью по функции δ, то

$$\left(\langle x, b - Ay_\delta(x) \rangle - \varphi\left(y_\delta(x)\right); b - Ay_\delta(x) \right)$$

будет $(\delta, 2L)$-моделью функции $f(x)$ в точке x относительно 2-нормы (см. начало § 3).

Обобщите этот результат на случай, когда

$$f(x)=\max_{y\in\hat{Q}}\left\{\langle x,h(y)\rangle-\varphi(y)\right\}\to\min_{x\in\mathbb{R}^n},$$

считая, что

$$\frac{1}{\mu}\max_{y\in\hat{Q},\,\|z\|_p\le 1}\left\|\nabla h(y)z\right\|_2^2<\infty,\ \frac{1}{\mu}\max_{y,u\in\hat{Q},\,\|z\|_p\le 1}\left\|\nabla^2 h(y)[z]u\right\|_2<1.$$

2) Задачу вида (4.7) можно решать с помощью *модифицированной функции Лагранжа* (*augmented Lagrangians*) [109]. В основе подхода лежит идея переписывания задачи (4.7) следующим образом ($\mu\ge 0$ – выбираемый параметр):

$$\varphi(y)+\frac{\mu}{2}\|Ay-b\|_2^2\to\min_{Ay=b,\ y\in\hat{Q}},$$

и стандартный переход к двойственной задаче:

$$f(x)=\max_{y\in\hat{Q}}\underbrace{\left\{\langle x,b-Ay\rangle-\varphi(y)-\frac{\mu}{2}\|Ay-b\|_2^2\right\}}_{\Psi(y,x)}\to\min_{x\in\mathbb{R}^n}.$$

Покажите, что если $y_\delta(x)$ – решение вспомогательной задачи максимизации в смысле (3.5):

$$\max_{y\in\hat{Q}}\left\langle\nabla_y\Psi(y_\delta(x),x),y-y_\delta(x)\right\rangle\le\delta,$$

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. [130, 140, 235]. ■

◊ Описанный во второй части упражнения 4.2 метод модифицированной функции Лагранжа лежит в основе одного из самых популярных алгоритмов распределенной оптимизации ADMM [115], см. пример 4.1. ◊

Замечание 4.3 (метод штрафных функций). Метод модифицированной функции Лагранжа тесно связан с методом *штрафных функций* [11, § 16, гл. 5]. Для полноты картины приведем 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:

$$f(x)+\frac{K}{2}\|g(x)\|_2^2 \to \min_x . \tag{4.31}$$

К задаче (4.31) можно прийти, например, *релаксировав* исходную постановку следующим образом [97]:

$$f(x) \to \min_{\frac{1}{2}\|g(x)\|_2^2 \le \frac{1}{2}\varepsilon^2} .$$

В таком случае $K := K(\varepsilon)$ можно понимать как множитель Лагранжа к ограничению

$$\frac{1}{2}\|g(x)\|_2^2 \le \frac{1}{2}\varepsilon^2 .$$

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, [11, § 17, гл. 5], [61, п. 5 § 2, гл. 8])

$$Kg(x^K) \xrightarrow[K\to\infty]{} \lambda \text{, т. е. } g(x^K) \simeq \frac{\lambda}{K},$$

$$f(x^K) - f(x_*) = \mathrm{O}\left(\frac{\|\lambda\|_2^2}{K}\right),$$

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 [29]. ■

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

$$\|\nabla f(x)\|_2 \le \|\nabla f^\mu(x)\|_2 + \mu\|x\|_2 ,$$

$$\langle x, \nabla f(x)\rangle \le \frac{L_\mu}{\mu}\left(f^\mu(x) - f^\mu(x_*^\mu)\right),\ L_\mu = L+\mu ,$$

$$\left\|x_*^\mu\right\|_2^2 \le \frac{2}{\mu}\left(f(0)-f(x_*)\right),$$

где для всех $x, y \in \mathbb{R}^n$ по постановке имеет место неравенство

$$\left\|\nabla f(y)-\nabla f(x)\right\|_2 \le L\left\|y-x\right\|_2.$$

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:

$$\left\langle x^N, \nabla f\left(x^N\right)\right\rangle \left(\le \frac{L_\mu}{\mu}\left(f^\mu\left(x^N\right)-f^\mu\left(x_*^\mu\right)\right)\right) \le \varepsilon,$$

$$\left\|\nabla f\left(x^N\right)\right\|_2 \left(\le \left\|\nabla f^\mu\left(x^N\right)\right\|_2 + \mu\left\|x^N\right\|_2\right) \le \tilde{\varepsilon}.$$

Учитывая, что $\left\|x^N\right\|_2 \le 2\left\|x_*^\mu\right\|_2$ (см. (1.12)), предложите способ выбора параметра регуляризации μ. Сопоставьте полученную таким образом оценку скорости сходимости с оценкой (4.16), учитывая, что $\left\|x_*^\mu\right\|_2 \le \left\|x_*\right\|_2$.

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 [16]. Из неравенства (1.8) имеем

$$f^\mu(x)-f^\mu\left(x_*^\mu\right) \ge \frac{\left\|\nabla f^\mu(x)\right\|_2^2}{2L_\mu} = \frac{\left\|\nabla f(x)+\mu x\right\|_2^2}{2L_\mu} \ge \frac{\mu\left\langle \nabla f(x), x\right\rangle}{L_\mu}.$$

Из неравенства (1.14) имеем

$$\frac{\mu}{2}\left\|x_\mu^*\right\|_2^2 = \frac{\mu}{2}\left\|0-x_\mu^*\right\|_2^2 \le f^\mu(0)-f^\mu\left(x_*^\mu\right) \le f(0)-f(x_*).$$

Последнее неравенство имеет место ввиду $f^\mu(0) = f(0)$ и

$$f^\mu\left(x_*^\mu\right) = f\left(x_*^\mu\right)+\frac{\mu}{2}\left\|x_*^\mu\right\|_2^2 \ge f(x_*)+\frac{\mu}{2}\left\|x_*^\mu\right\|_2^2 \ge f(x_*).$$

В случае наличия у $f(x)$ представления (4.8) имеет место также оценка

$$f(0)-f(x_*) \le \min_{Ay=b,\ y\in\tilde{Q}} \varphi(y) - \min_{y\in\tilde{Q}} \varphi(y).$$

Заметим, что для оценки $\left\|x_*^\mu\right\|_2$ сверху можно было бы также использовать неравенство $\left\|x_*^\mu\right\|_2 \le \left\|x_*\right\|_2$ и (4.16). ■

Упражнение 4.5. Обобщите рассуждения § 4 в случае, когда в задаче (4.22) вместо неравенства $Ay \le b$ рассматриваются общие выпуклые ограничения: $Ay = b$, $h(y) \le 0$.

Указание. См. [234]. ■

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) задается небольшим числом сепарабельных выпуклых неравенств вида $\sum_{i=1}^{n} h_i^j(x_i) \le 0$, $j = 1,...,m$, то задачи вида (2.6), (2.29) и при определенных условиях (3.3) могут быть решены (в смысле (3.4)) за время $O\left(nm^2 \ln^2(n/\varepsilon)\right)$.

Указание. Характерный пример получения такой оценки разбирается в [22] (см. также [223, п. 5.3.3]). В основе подхода – решение малоразмерной двойственной задачи каким-нибудь прямодвойственным быстро (линейно) сходящимся методом типа метода эллипсоидов [133, 226], 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 расчета 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, то все «огрубления», накопленные по ходу описанных рассуждений, соберутся под логарифмом и испортят лишь мультипликативную константу в итоговой оценке. Отметим также, что в (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 $\|x\|_p^p \le y^{p/2}$, где $p/2 < 1$. ■

Упражнение 4.7 («нащупывание» цен по Вальрасу и централизованная распределенная оптимизация [14]). Пусть руководство города владеет n пекарнями. Затраты i-й пекарни на выпечку x_i тонн хлеба в день равны $f_i(x_i)$ – сильно выпуклые возрастающие функции. Задача руководства: производить не меньше C тонн хлеба в день (C – объём спроса на хлеб в день со стороны населения города) так, чтобы суммарные затраты всех пекарен были бы минимальны. Формально задача может быть поставлена следующим образом:

$$\sum_{i=1}^{n} f_i(x_i) \to \min_{\substack{\sum_{i=1}^{n} x_i \geq C \\ x_i \geq 0,\ i=1,\dots,n}}. \tag{4.33}$$

Обозначим решение этой задачи $x^* = \left\{x_i^*\right\}_{i=1}^n$.

1) Предположим теперь, что у пекарен есть собственники, которые продают хлеб руководству города, распределяющего этот хлеб среди населения, по цене p^k в k-й день. Таким образом, собственники решают задачи:

$$x_i\left(p^k\right) = \arg\max_{x_i \geq 0} \overbrace{\left\{\underbrace{p^k x_i}_{\text{выручка}} - \underbrace{f_i(x_i)}_{\text{затраты}}\right\}}^{\text{прибыль}},\ i=1,\dots,n. \tag{4.34}$$

Руководство города действует по следующему правилу: в каждый день k у руководства есть представление о том, в каком отрезке лежит равновесная цена $\left[p_{\min}^k, p_{\max}^k\right]$. Выставив цену $p^k = \frac{1}{2}\left(p_{\min}^k + p_{\max}^k\right)$, руководство собирает следующую информацию с пекарен: $\sum_{i=1}^{n} x_i\left(p^k\right)$. Далее

$$\left[p_{\min}^{k+1}, p_{\max}^{k+1}\right] = \left[p_{\min}^k, \frac{1}{2}\left(p_{\min}^k + p_{\max}^k\right)\right], \text{ если } \sum_{i=1}^{n} x_i\left(p^k\right) > C;$$

$$\left[p_{\min}^{k+1}, p_{\max}^{k+1}\right] = \left[\frac{1}{2}\left(p_{\min}^k + p_{\max}^k\right), p_{\max}^k\right], \text{ если } \sum_{i=1}^{n} x_i\left(p^k\right) \leq C.$$

Покажите, что

$$x_i\left(p^k\right) \xrightarrow[k\to\infty]{} x_i^*.$$

Попробуйте оценить скорость сходимости. Предложите способ оценки $\left[p_{\min}^0, p_{\max}^0\right]$.

2) Переписав задачу (4.33) следующим (равносильным) образом:

$$\sum_{i=1}^{n} f_i(x_i) \to \min_{\substack{x_i \ge y_i,\, i=1,\dots,n \\ \sum_{i=1}^{n} y_i \ge C \\ x_i, y_i \ge 0,\, i=1,\dots,n}},$$

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 представление о производственных процессах на всех пекарнях. С одной стороны, описанный выше процесс можно понимать как процесс нащупывания равновесной цены [60, гл. 10]. С другой стороны, этот процесс можно понимать как распределенный централизованный алгоритм решения задачи выпуклой оптимизации (4.33) [221]: задача хранится на n рабочих узлах/пекарнях (slave nodes), взаимодействие которых осуществляется через центр/руководство (master node). На каждом узле осуществляется работа только со своей частью задачи (решаются вспомогательные задачи (4.34)), коммуникация осуществляется, как показано на рис. 8.

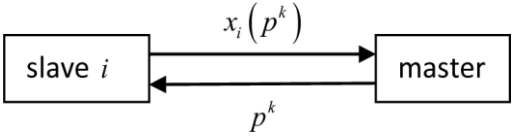

Рис. 8

1) Для решения задачи нужно построить двойственную задачу (с точностью до знака):

$$\psi(p) = \sum_{i=1}^{n} \left(p x_i(p) - f_i(x_i(p)) \right) - Cp \to \min_{p \ge 0}$$

и заметить, что

$$\psi'(p) \overset{def}{=} \frac{d\psi(p)}{dp} = \sum_{i=1}^{n} x_i(p) - C\,.$$

Из условий задачи $p^0_{\min} \geq 0$, 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

$$\begin{aligned} \psi(p) &:= \psi(p_1,\dots,p_n) = \\ &= \sum_{i=1}^{n}\left(p_i x_i(p_i) - f_i\left(x_i(p_i)\right)\right) - C\min_{i=1,\dots,n} p_i \to \min_{p=(p_1,\dots,p_n)\in\mathbb{R}^n_+}, \end{aligned} \tag{4.35}$$

где p_i – множитель Лагранжа к ограничению $x_i \geq y_i$. Тогда

$$\partial\psi(p) = \begin{pmatrix} x_1(p) \\ \vdots \\ x_n(p) \end{pmatrix} - C\begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix},\ \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix} \in S_n(1),\ \lambda_i > 0 \ \Rightarrow\ i \in \operatorname{Arg}\min_{l=1,\dots,n} p_l .$$

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 [14, 234] или *сходящийся субградиентный метод* Нестерова–Шихмана, который будет иметь в данном контексте вполне естественную интерпретацию [240, 241]. Отличие этого метода от близкого ему субградиентного метода [14], 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 [14, 240, 241], сохранив возможность содержательной интерпретации, если смотреть на слагаемое $C\min\limits_{i=1,\dots,n} p_i$ в задаче (4.35) как на композитный член (см. пример 3.1). ■

Упражнение 4.8 (децентрализованная распределенная оптимизация на меняющихся со временем графах [258]). Покажите, что если в примере 4.1 дополнительно предположить, что $\varphi_i''(y) \leq L_\varphi$, то «двойственная» функция $f(x)$ (4.20) будет сильно выпуклой в 2-норме с константой $\mu_f = \tilde{\sigma}_{\min}(W)/L_\varphi$ на $\operatorname{Ker}\left(W^T\right)^\perp = \operatorname{Ker}(W)^\perp$ (при редуцированном подходе $\mu_f = \tilde{\sigma}_{\min}\left(\sqrt{W}\right)/L_\varphi$ и на $\operatorname{Ker}\left(\sqrt{W}\right)^\perp$). Найдите оценку скорости

сходимости метода из примера 4.1, если ребра графа $G = \langle V, E \rangle$ со временем как-то меняются, при этом все время сохраняется связность графа.

Указание. См. [147, утверждение 2.1], [235, теорема 1], [190, теорема 6]. 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 [47, п. 1.4, 2.2] отчасти объясняет отмеченную ранее асимметрию в возможности адаптивной настройки методов на неизвестную константу Липшица градиента и на отсутствие такой возможности для константы сильной выпуклости. Во всяком случае, пока не придумали, как можно было в общем случае адаптивно осуществлять такую настройку без серьезных дополнительных усилий, см. указание к упражнению 1.3. ■

§ 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 методу Ю.Е. Нестерова [238] (см. также [5, 15, 24, 72, 163, 285]), будет сделан, пожалуй, самый важный шаг во всем описанном выше подходе – согласованы формулы (2.5), (2.26), (2.27). Как уже ранее отмечалось, 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 использовался шаг $h = 1/L$, где константа L либо была задана по условию, либо определялась согласно (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

$$L^{k+1} = L^k/2,$$

While True Do

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

If $\left\{ f\left(x^{k+1}\right) \le f\left(x^k\right) + \left\langle \nabla f\left(x^k\right), x^{k+1} - x^k \right\rangle + L^{k+1} V\left(x^{k+1}, x^k\right) + \dfrac{\varepsilon}{2} \right\}$

Перейти на следующую итерацию: $k \to k+1$

Else

$$L^{k+1} := 2L^{k+1}.$$

Для такого метода формула (4.1) перепишется следующим образом:

$$\begin{aligned} \frac{1}{L^{k+1}} f\left(x^{k+1}\right) \le \frac{1}{L^{k+1}} \left\{ f\left(x^k\right) + \left\langle \nabla f\left(x^k\right), x - x^k \right\rangle \right\} + \\ + V\left(x, x^k\right) - V\left(x, x^{k+1}\right) + \frac{\varepsilon}{2L^{k+1}}, \end{aligned} \tag{5.1}$$

где константа L^{k+1} подбирается в (5.1) согласно описанной выше процедуре оптимально (с точностью до множителя 2) из соотношения (2.26), в котором $\delta = \varepsilon/2$:

$$f\left(x^{k+1}\right) \le f\left(x^k\right) + \left\langle \nabla f\left(x^k\right), x^{k+1} - x^k \right\rangle + \frac{L^{k+1}}{2}\left\| x^{k+1} - x^k \right\|^2 + \frac{\varepsilon}{2}.$$

То есть автоматически происходит подбор на рассматриваемом отрезке $\left[x^k, x^{k+1}\right]$ двух параметров ν и L_ν в (2.27):

$$\left\| \nabla f(y) - \nabla f(x) \right\|_* \le L_\nu \left\| y - x \right\|^\nu, \nu \in [0,1],\ L_0 < \infty$$

так, чтобы (2.26) выполнилось. Подчеркнем, что не мы сами решаем задачу подбора ν и L_ν – это делает метод за счет описанной процедуры. Свойства гладкости функции $f(x)$ на отрезке $\left[x^k, x^{k+1}\right]$ характеризуются континуальным набором чисел $\left\{L_\nu\right\}_{\nu\in[0,1]}$, часть из которых может равняться бесконечности. Мы можем ничего не знать о $\left\{L_\nu\right\}_{\nu\in[0,1]}$ – этого и не требуется при универсальном подходе. Тем не менее описанная выше процедура гарантирует, что метод подберет такое $\nu \in [0,1]$, что соответствующая этому ν константа Гёльдера L_ν порождает (по формуле (2.27) с $\delta = \varepsilon/2$) на отрезке $\left[x^k, x^{k+1}\right]$ минимально возможную (с точностью до множителя не больше 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*[22]

$$N = \inf_{\nu\in[0,1]} \left(\frac{2L_\nu R^{1+\nu}}{\varepsilon} \right)^{\frac{2}{1+\nu}} \tag{5.2}$$

итераций имеет место следующая оценка:

[22] Причем в (5.2) константы L_ν выбираются не худшие по всему пространству, а «средние» на траектории метода. При получении оценки (5.2) предполагалось, что L^0 выбиралось 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 имеет место, то это приведет лишь к небольшому ухудшению числового множителя в оценке (5.2).

$$f\left(\overline{x}^N\right)-f\left(x_*\right)\le f\left(\overline{x}^N\right)-$$
$$-\frac{1}{\sum\limits_{k=0}^{N-1}1/L^{k+1}}\min\limits_{x\in B_{R,Q}\left(x^0\right)}\left\{\sum_{k=0}^{N-1}\frac{1}{L^{k+1}}\left[f\left(x^k\right)+\left\langle\nabla f\left(x^k\right),x-x^k\right\rangle\right]\right\}\le\varepsilon,$$

где (*в данном случае*)

$$\overline{x}^N=\frac{1}{\sum\limits_{k=1}^{N}1/L^k}\sum_{k=1}^{N}\frac{x^k}{L^k},$$

$R^2=V\left(x_*,x^0\right)$. *Если решение* x_* *не единственно, то оценка* (5.2) *справедлива и для того решения* x_*, *которое доставляет минимум* R^2.

Еще раз подчеркнем, что задачу минимизации, возникающую в оценке (5.2), нам не нужно решать, равно как и знать хоть что-то о гладкости $f(x)$, кроме того, что $L_0<\infty$, чтобы универсальный метод сходился. В негладком случае (когда только $L_0<\infty$) имеем $N=4L_0^2R^2/\varepsilon^2$, что с точностью до множителя соответствует нижней оценке, см. (2.37).

Рассуждая аналогично [238], несложно показать, что описанный выше универсальный метод на каждой итерации запрашивает один раз $\nabla f\left(x^k\right)$ и в среднем (по итерациям) около трех раз значение функции $f(x)$. Действительно, среднее число вычислений значения функции $f(x)$ есть

$$\frac{1}{N}\sum_{k=0}^{N-1}\left(2+\log_2\left(L^{k+1}/\left(L^k/2\right)\right)\right)=3+\frac{1}{N}\log_2\left(L^N/L^0\right).$$

Замечание 5.1 (универсальный метод для вариационных неравенств и седловых задач [19]).[23] Аналогом градиентного метода для *вариационных неравенств* (ВН) и седловых задач является *экстраградиентный метод* Г. М. Корпелевич [11, § 15, гл. 5], [46]. Крупным специалистом, активно работающим долгие годы в этом направлении в России, является А. С. Антипин [98]. Развитие этих методов можно проследить по его работам. Далее рассмотрим один современный вариант экстраградиентного метода, а именно *проксимальный зеркальный метод*

[23] В концепции модели функции из примера 1 § 3 на базе этого замечания можно получить универсальный вариант популярного сейчас *метода Chambolle–Pock'a* [126, 214].

А. С. Немировского [224]. 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

$$\langle g(y)-g(x),y-z\rangle \le LV(y,x)+LV(y,z)+\delta .$$

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:

$$\frac{1}{N}\sum_{k=1}^{N}\langle g(y^k),y^k-x\rangle \le \frac{LV(x,x^0)-LV(x,x^N)}{N}+\delta. \qquad (5.3)$$

С помощью текста, написанного в конце п. 4.6 [119], несложно построить универсальный вариант такого метода (описывается *k*-я итерация):

Универсальный проксимальный зеркальный метод

$$L^{k+1}=L^k/2,$$

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 $\left\{\langle g(y^{k+1})-g(x^k),y^{k+1}-x^{k+1}\rangle \le L^{k+1}V(y^{k+1},x^k)+L^{k+1}V(y^{k+1},x^{k+1})+\frac{\varepsilon}{2}\right\}$

Перейти на следующую итерацию: $k\to k+1$

Else

$$L^{k+1}:=2L^{k+1}.$$

Если, подобно (2.27), векторное поле $g(x)$ удовлетворяет условию

$$\|g(y)-g(x)\|_* \le L_\nu \|y-x\|^\nu,\ \nu\in[0,1],\ x,y\in Q,\ L_0<\infty,$$

то, используя неравенство (верное для любых $a,b,L_\nu,\delta>0$, $\nu\in[0,1]$)

$$L_\nu a^\nu b \le L_\nu\cdot\left(\frac{L_\nu}{\delta}\right)^{\frac{1-\nu}{1+\nu}}\left(\frac{a^2}{2}+\frac{b^2}{2}\right)+\delta$$

с $\delta = \varepsilon/2$, можно получить (см. [19]), что для достижения

$$\frac{1}{\sum\limits_{k=1}^{N} 1/L^k} \max_{x\in Q}\left\{\sum_{k=1}^{N}\frac{1}{L^k}\left\langle g\left(y^k\right), y^k - x\right\rangle\right\} \le \varepsilon \tag{5.4}$$

достаточно (следует сравнить с (5.2))

$$N = \inf_{\nu\in[0,1]}\left(\frac{2L_\nu R^{1+\nu}}{\varepsilon}\right)^{\frac{2}{1+\nu}} \tag{5.5}$$

итераций универсального проксимального зеркального метода. Здесь $R^2 = \max\limits_{x\in Q} V\left(x, x^0\right)$. При этом среднее число вычислений значений векторного поля $g(x)$ на одной итерации приближенно равно трем. Оценка (5.5) с точностью до числового множителя оптимальна для ВН и для седловых задач. К сожалению, точной ссылки на обоснование оптимальности не удалось найти, однако различные частные случаи могут быть сведены к разобранному в работах [52, 177, 222].

◊ Следует сравнить (5.4) при $g(x) = \nabla f(x)$ с (2.25), (3.4). Если $g(x) = \left(\nabla_u f(u,w), -\nabla_w f(u,w)\right)$, $x = (u,w)$, $Q = Q_u \otimes Q_w$, где функции $f(u,w)$ выпуклая по u и вогнутая по w, то из (5.4) следует, что

$$0 \le \max_{w\in Q_w} f\left(\bar{u}^N, w\right) - \min_{u\in Q_u} f\left(u, \bar{w}^N\right) \le \varepsilon.$$

Отметим при этом, что для седловой точки (u_*, w_*) имеем

$$\max_{w\in Q_w} f(u_*, w) = \min_{u\in Q_u} f(u, w_*). \text{ ◊}$$

Заметим, что для *монотонных вариационных неравенств*[24]

$$\left\langle g(y) - g(x), y - x\right\rangle \ge 0,\ x, y \in Q,$$

имеем

$$\left\langle g(x), y^k - x\right\rangle = \left\langle g\left(y^k\right), y^k - x\right\rangle + \underbrace{\left\langle g(x) - g\left(y^k\right), y^k - x\right\rangle}_{\le 0} \le \left\langle g\left(y^k\right), y^k - x\right\rangle.$$

В этой связи формулу (5.4) можно переписать как[25]

[24] В случае $g(x) = \nabla f(x)$ это условие соответствует выпуклости $f(x)$.

$$\max_{x\in Q}\left\langle g(x), \bar{y}^N - x\right\rangle \le \varepsilon\,, \tag{5.6}$$

где

$$\bar{y}^N = \frac{1}{\sum\limits_{k=1}^{N} 1/L^k}\sum_{k=1}^{N}\frac{y^k}{L^k}\,.$$

В литературе обычно используют именно этот критерий качества решения монотонных ВН, см., например, [53, гл. 3], [224].

Подобно слабой квазивыпуклости из замечания 2.1 можно ослабить условия монотонности ВН, во многом сохранив результаты [135]. ■

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, [5]. 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 (2.34)). Отказавшись от универсальности, можно избавиться и от этого ограничения, используя *метод двойственных усреднений* Ю. Е. Нестерова [234].[26] 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 [71]. Отметим тем не менее, что здесь речь идет только о простоте выводов, но не о потенциальных возможностях вывода. В-третьих, привязка к (1.32) позволяет переносить результаты, полученные непосредственно для негладких выпуклых задач, т. е. без универсализации, на онлайн-постановки, в том числе стохастические и сильно вы-

---

[25] Собственно под решением вариационного неравенства обычно понимают следующую задачу: найти такую точку $x_*\in Q$, что для всех $x\in Q$ $\left\langle g(x), x_* - x\right\rangle \le 0$, т. е. $\max\limits_{x\in Q}\left\langle g(x), x_* - x\right\rangle \le 0$. Отсюда становится ясным смысл условия (5.6). Подробнее о ВН см., например, [6, 26], [36, Ч. 3], [53, гл. 3].

[26] Этот метод близок другому популярному методу решения негладких задач оптимизации – *методу зеркального спуска* А. С. Немировского, см., например, [92], [223, гл. 5].

пуклые[27], см., например, [180]. В-четвертых, 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, [61, п. 5 § 1 и п. 3 § 3, гл. 5]), в которой отсутствуют правые неравенства в (2.3), (3.1).[28] В-пятых, при перенесении универсальных методов на стохастические постановки задач [24], 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 [15]. Связано это, прежде всего, с тем, что при универсальном подходе необходимо рассчитывать значение функции, что может быть намного дороже расчета значения ее стохастического градиента. Вот простой пример [186]:

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

где $A \succ 0$ – плотно заполненная неотрицательно определенная матрица $n \times n$, $x \in S_n(1)$. Несмещенный стохастический градиент этой функции

$$\nabla_x f(x, j) = A^j, \text{ где } P(j = i) = x_i,\ i = 1, \ldots, n.$$

Ясно, что

$$E_j\left[\nabla_x f(x, j)\right] = Ax = \nabla f(x).$$

Также ясно, что на подсчет $f(x)$ уходит время $\mathrm{O}(n^2)$, в то время как на подсчет $\nabla_x f(x, j)$ – время $\mathrm{O}(n)$. 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. ■

Вернемся к оценке (5.1). Попробуем с помощью этой оценки и *техники рестартов* (упражнение 2.3) распространить описанный выше

---

[27] 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 ее корректировка в части $L_0^2/(\mu N) \to L_0^2 \ln N/(\mu N)$. Такая нижняя оценка уже будет достижима [181].

[28] Работая с δ-субградиентами где $\delta = \mathrm{O}(\varepsilon/N(\varepsilon))$, вместо настоящих субградиентов можно получать, например, оценки из указания к упражнению 1.4, изначально установленные для работы с настоящими субградиентами [52, 97], см. также упражнение 5.5.

универсальный градиентный спуск на сильно выпуклые задачи. Ввиду теоремы 1.1 отметим, что существуют и другие способы того, как это можно сделать. Однако выбранный здесь способ представляется наиболее удачным в методическом плане своей общеприменимостью.

Итак, подобно выводу (4.5) из (4.1), из (5.1) можно получить

$$f\left(\overline{x}^N\right)-f\left(x_*\right)\le\frac{\overline{L}_N V\left(x_*,x^0\right)}{2N}+\frac{\varepsilon}{2}, \tag{5.7}$$

где

$$\overline{L}_N=\frac{N}{\sum\limits_{k=1}^{N}1/L^k},\ \overline{x}^N=\frac{\overline{L}_N}{N}\sum_{k=1}^{N}\frac{x^k}{L^k}.$$

Пусть $f\left(x\right)$ – μ-сильно выпуклая функция в норме $\|\ \|$, согласованной с дивергенцией Брэгмана $V\left(y,x\right)$ (см. § 2, в частности, формулу (2.32)). Пусть $d\left(x-x^0\right)\le C_n\left\|x-x^0\right\|^2$ (можно считать $C_n=\mathrm{O}\left(\ln n\right)$, см. п. 2) упражнения 2.3). Тогда из (1.14) и (5.7) следует, что

$$\frac{\mu}{2}\left\|\overline{x}^{N_1}-x_*\right\|^2\le f\left(\overline{x}^{N_1}\right)-f\left(x_*\right)\le\frac{\overline{L}_{N_1}V\left(x_*,x^0\right)}{2N_1}+\frac{\varepsilon}{2}\le\frac{C_n\overline{L}_{N_1}\left\|x^0-x_*\right\|^2}{2N_1}+\frac{\varepsilon}{2}.$$

Далее используется схема рассуждений, аналогичная той, что была изложена в указании к упражнению 2.3. А именно, из соотношения

$$\frac{\mu}{2}\left\|\overline{x}^{N_1}-x_*\right\|^2\le\frac{C_n\overline{L}_{N_1}\left\|x^0-x_*\right\|^2}{2N_1}+\frac{\varepsilon}{2} \tag{5.8}$$

выбираем наименьшее такое N_1, при котором

$$\left\|\overline{x}^{N_1}-x_*\right\|^2\le\frac{1}{2}\left\|x^0-x_*\right\|^2. \tag{5.9}$$

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[29] из оценки (5.2) получится оценка вида [17, 22, 157]:

---

[29] На $(k+1)$-м рестарте следует выбирать точку старта как $x^0=\overline{x}^{N_k}$ – среднее арифметическое точек, полученных по ходу работы метода на k-м рестарте, а $d\left(x\right):=d\left(x-x^0\right)=d\left(x-\overline{x}^{N_k}\right)$ [24, 186, 188]. Натуральный логарифм в (5.10)

$$N = \mathrm{O}\Bigg(\underbrace{C_n \inf_{\nu\in[0,1]} \left(\frac{L_\nu^2}{\mu^{1+\nu}\varepsilon^{1-\nu}} \right)^{\frac{1}{1+\nu}}}_{\text{число итераций на одном рестарте}} \underbrace{\left\lceil \ln\left(\frac{\mu \left\| x_* - x^0 \right\|^2}{\varepsilon} \right) \right\rceil}_{\text{число рестартов}} \Bigg), \qquad (5.10)$$

где 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 (5.10) с точностью до C_n, получим оценку (5.2).

Основная проблема с реализацией описанного выше подхода – это явное использование в нем, как правило, неизвестного параметра μ. Довольно естественный способ борьбы с неизвестностью параметра μ состоит в рестартах по невязке $f\left(x^N\right) - f\left(x_*\right)$. Если $f\left(x_*\right)$ известно, то рестарты можно делать, контролируя эту невязку по функции [157, 161, 259].

Замечание 5.3 (контроль нормы градиента). Для неускоренного (универсального) градиентного спуска также можно использовать рестарты по норме градиента (градиентного отображения – см. упражнение 3.4).[30] Ограничимся далее для наглядности задачей безусловной выпуклой оптимизации (1.1) в условиях (1.4). Тогда для метода (1.22)

$$x^{k+1} = x^k - \frac{1}{L}\nabla f\left(x^k\right)$$

имеет место следующая не улучшаемая для данного метода оценка [275] (следует сравнить с оценкой (1.23) – оптимальной для класса гладких невыпуклых задач):

$$\left\| \nabla f\left(x^N\right) \right\|_2 \le \frac{LR}{N+1},$$

выбран потому, что именно так наиболее часто выбирают в данном контексте в литературе.

[30] Следует сопоставить 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).

где $R = \left\| x^0 - x_* \right\|_2$. 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) [275]. Ситуация меняется для ускоренных методов, см. указание к упражнению 1.3. Для ускоренных (быстрых, моментных) градиентных спусков на данный момент удается получить (см. (1.8)) лишь оценки вида [194, 230, 275]:

$$\left\| \nabla f\left(x^N\right) \right\|_2 \le \sqrt{2L \cdot \left(f\left(x^N\right) - f\left(x_*\right) \right)} \le \frac{2LR}{N+1}, \ \min_{k=0,\dots,N} \left\| \nabla f\left(x^k\right) \right\|_2 \le \frac{8LR}{N^{3/2}},$$

которые не улучшаемы больше, чем на числовой множитель для рассмотренных на данный момент классов (ускоренных) методов первого порядка [275]. По-прежнему открытым остается вопрос, можно ли добиться для какого-нибудь из методов первого порядка сходимости вида $\left\| \nabla f\left(x^N\right) \right\|_2 \sim LR/N^2$ [230, 275]. 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 [24, 53, 54, 119, 141], то с точностью до $\ln N$ получается желаемая оценка.[31] К сожалению, для всех известные 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 ее 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. ■

Для задач безусловной выпуклой оптимизации замечание 5.3 также дает возможность контролировать только норму градиента в качестве критерия останова неускоренного градиентного спуска и его универсального варианта.

Отметим также, что недавно была обнаружена возможность (не связанная с использованием прямодвойственности) избавления от знания

[31] Упражнения 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 приближенному решению двойственной задачи решения прямой задачи.

значения параметра μ и в рестартах для ускоренных вариантов градиентных методов [159].

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 $\delta = \mathrm{O}(\varepsilon)$ и $\tilde{\delta} = \mathrm{O}(\varepsilon)$ (см. обозначения § 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. Оптимизируемая (целевая) функция $f(x)$, как и раньше, предполагается выпуклой на выпуклом множестве Q. Условие гладкости и сильной выпуклости $f(x)$ при наличии шума следует понимать следующим образом (3.1) [141, 151]:

$$\frac{\mu}{2}\|y-x\|^2 \le f(y) - \left(f_\delta(x) + \psi_\delta(y,x)\right) \le \frac{L}{2}\|y-x\|^2 + \delta.$$

Впрочем, используя технику рестартов, можно исходить и из обычной концепции модели функции (3.1), предполагая сильную выпуклость у модели функции (3.1) $\psi_\delta(y,x)$ как функции y или (еще более общий случай) предполагая сильную выпуклость только у $f(x)$ [24, 157]. ■

Упражнение 5.4. С помощью техники рестартов (см. упражнение 2.3) и формулы (5.3) (или как-то по-другому) попробуйте перенести результаты замечания 5.1 на *сильно монотонные вариационные неравенства* и сильно выпукло/вогнутые седловые задачи [53, гл. 3]. Отметим одно затрудняющее такой перенос обстоятельство: в формуле (5.5) используется $R^2 = \max\limits_{x\in Q} V(x, x^0)$, что не дает 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. См., например, [66]. ■

Упражнение 5.5. Используя принцип множителей Лагранжа [116, гл. 5], распространите проксимальный зеркальный метод из замечания 5.1 на общие задачи выпуклого программирования [47, п. 2], предварительно компактифицировав двойственные переменные (см., например, упражне-

ние 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 [133, 226]. Для этого потребуется число итераций (см. указание к упражнению 1.4) $N_{ellips}(\varepsilon) \sim \ln\left(\varepsilon^{-1}\right)$. Причем на каждой итерации вместо настоящего субградиента можно использовать δ-субградиент (см. замечание 5.2), где

$$\delta \sim \varepsilon / N_{ellips}(\varepsilon),$$

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 [61, 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. Причем решение задачи доставляет в этом неравенстве 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. концовку § 5 и [22].

Отметим, что если в условии задачи имеется несколько сильно выпуклых ограничений вида неравенств $h_1(x) \le 0$, …, $h_m(x) \le 0$, то их можно заменить скалярным негладким сильно выпуклым ограничением: $h(x) = \max\{h_1(x), ..., h_m(x)\} \le 0$ – см. [1], [61, п. 3 § 3, гл. 10], [237, 242] и замечание 3.1. ■

Упражнение 5.7. Предложите способ распространения проксимального зеркального метода из замечания 5.1 на бесконечномерные задачи, например, дифференциальные игры [158].

Упражнение 5.8. Определите, какие из результатов, описанных выше (во всем 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. В случае замечания 5.1 частичное ускорение при некоторых дополнительных предположениях оказывается возможным [129, 130, 147]. В случае упражнения 4.8 возникают дополнительные сложности при перенесении. Тем не менее на базе результатов работ [103, 159], по-видимому, можно осуществить желаемое перенесение с некоторыми оговорками. В связи с упоминанием в таком контексте упражнения 4.8 отметим также, что в примере 4.1 и в упражнении 4.8 можно отказаться и от свойства неориентированности коммуникационного графа / симметричности матрицы W, сохраняя при этом его связность. В этом случае, используя другую технику, можно получить результаты, похожие на те, что были приведены в пособии, см., например, [220] 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 [281].

Отметим, что результаты, связанные с относительной гладкостью из § 3, по-видимому, переносятся на ускоренные методы лишь при дополнительных обременительных предположениях на дивергенцию Брэгмана [178, 213]. Результаты, касающиеся α-слабой квазивыпуклости переносятся на ускоренные методы, если дополнительно допускается вспомогательная маломерная оптимизация на каждом шаге [175]. Пока только при $\alpha = 1$ удалось избавиться от этого ограничения [71, замечание 7]. ■

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 $N \le n$ на скорость решения такой задачи (см. также упражнение 1.3 и замечание 1.6) имеет вид

$$\|Ax^N - b\|_2^2 \ge \frac{L_x R_x^2}{2(2N+1)^2},$$

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 наименьшей 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 [116, гл. 5]:

$$\min_{Ax=b}\frac{1}{2}\|x\|_2^2 = \min_x \max_\lambda \left\{\frac{1}{2}\|x\|_2^2 + \langle b - Ax, \lambda\rangle\right\} = \max_\lambda \min_x \left\{\frac{1}{2}\|x\|_2^2 + \langle b - Ax, \lambda\rangle\right\} =$$

$$= \max_\lambda \left\{\langle b, \lambda\rangle - \frac{1}{2}\left\|A^T\lambda\right\|_2^2\right\}.$$

С помощью теоремы 4.1 и упражнений 1.3, 5.8 покажите, что после N итераций ускоренного градиентного метода, примененного к двойственной задаче, можно восстановить решение исходной задачи $\breve{x}^N$ со следующей точностью:

$$\left\|A\breve{x}^N - b\right\|_2 \le \frac{8L_\lambda R_\lambda}{N^2},$$

где $L_\lambda = \sigma_{\max}\left(A^T\right) = \sigma_{\max}(A)$, $R_\lambda = \|\lambda_*\|_2$. 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. [2, 18, 100, 161, 255]. Заметим, что поскольку система $Ax = b$ совместна, то по *теореме Фредгольма* [36, 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 x, что $Ax = b$, поэтому для всех λ имеет место: $\langle Ax, \lambda\rangle = \langle b, \lambda\rangle$. Следовательно, $\langle x, A^T\lambda\rangle = \langle b, \lambda\rangle$. Предположив, что существует такой λ, что $A^T\lambda = 0$ и $\langle b, \lambda\rangle > 0$, придем 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 линейной алгебры, вычислительной математики, но и численных методов оптимизации [117]. Напомним также, что класс сложности задач выпуклой 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 мат-

рицы) возникает на каждом шаге метода Ньютона (см. приложение). В свою очередь задачи квадратичной оптимизации получаются из $Ax = b$, обычно, либо как указано в упражнении 5.9, либо (в случае симметричности матрицы A) согласно (1.30). Упражнения 1.6, 5.9 (см. также [15, 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 спектральные свойства, также можно решать (используя рандомизированные методы) системы линейных уравнений за время, пропорциональное (с точностью до степеней логарифмических множителей, зависящих от желаемой точности) числу ненулевых элементов в матрице A [15, гл. 4], [131, 192, 256, 270]. ◊

Приложение. Обзор современного состояния развития численных методов выпуклой оптимизации

Описанные в § 2 – § 5 конструкции переносятся на ускоренные (быстрые, моментные) градиентные методы [72], например, на *метод подобных треугольников*[32] [92], см. также упражнения 3.7, 5.8. Причем 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 [17, 22, 72, 157, 238]:

$$N(\varepsilon) = \mathrm{O}\left(\inf_{\nu\in[0,1]} \left(\frac{L_\nu R^{1+\nu}}{\varepsilon} \right)^{\frac{2}{1+\nu}} \right) \to \mathrm{O}\left(\inf_{\nu\in[0,1]} \left(\frac{L_\nu R^{1+\nu}}{\varepsilon} \right)^{\frac{2}{1+3\nu}} \right),$$

$$N(\varepsilon) = \mathrm{O}\Bigg(\underbrace{C_n \inf_{\nu\in[0,1]} \left(\frac{L_\nu^2}{\mu^{1+\nu}\varepsilon^{1-\nu}} \right)^{\frac{1}{1+\nu}}}_{\text{число итераций на одном рестарте}} \underbrace{\left\lceil \ln\left(\frac{\mu\left\|x_* - x^0\right\|^2}{\varepsilon} \right) \right\rceil}_{\text{число рестартов}} \Bigg) \to$$

$$\to \mathrm{O}\Bigg(\underbrace{C_n \inf_{\nu\in[0,1]} \left(\frac{L_\nu^2}{\mu^{1+\nu}\varepsilon^{1-\nu}} \right)^{\frac{1}{1+3\nu}}}_{\text{число итераций на одном рестарте}} \underbrace{\left\lceil \ln\left(\frac{\mu\left\|x_* - x^0\right\|^2}{\varepsilon} \right) \right\rceil}_{\text{число рестартов}} \Bigg),$$

$$\delta = \mathrm{O}(\varepsilon) \to \tilde{\mathrm{O}}\left(\frac{\varepsilon}{N(\varepsilon)} \right).$$

[32] Заметим, что у метода линейного каплинга (МЛК) [92] за счёт наличия двух проектирований на каждой итерации имеются некоторые дополнительные свойства (по сравнению с методом подобных треугольников, у которого одно проектирование), обнаруженные недавно [13, 152, 153, 176]. 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 [92] Grad-шаг лучше заменить на Mirr-шаг, чтобы в случае неевклидовой прокс-структуры гарантировать равномерную ограниченность последовательности, генерируемой методом [2]. Другими словами, (1.35), (1.36) стоит заменять на (2.29) с соответствующим выбором размеров шагов.

Данные оценки являются неулучшаемыми (оптимальными) [138, 177, 222].

◊ В свою очередь, эти оценки можно обобщить на так называемые *промежуточные методы* [138, гл. 6], [151], методы, которые представляют собой выпуклые комбинации неускоренного и ускоренного градиентного метода: в выписанных формулах вместо $\left[\frac{1}{1+\nu};\frac{1}{1+3\nu}\right]$ стоит писать $\frac{1}{1+\nu+2p\nu}$, при этом $\delta=\tilde{\mathrm{O}}\left(\varepsilon/N(\varepsilon)^{p}\right)$, $p\in[0,1]$. Такого рода обобщения могут потребоваться, например, при решении задач оптимизации в гильбертовых пространствах [161]. Детали и дальнейшее обобщение на случай неточного проектирования (см. (3.3) и упражнение 3.7) имеется в работе [157]. ◊

Для большей наглядности далее (если не оговорено противного) рассматривается задача выпуклой безусловной оптимизации:

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

В качестве нормы выбирается 2-норма. В качестве прокс-функции: $d(x)=\frac{1}{2}\|x\|_2^2$, см. § 2.

◊ Тем не менее стоит отметить, что все 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 [13, 18, 20, 152, 153, 243].

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 [15, 18, 153, 156]. ◊

Изложенные выше результаты (в том числе и в ускоренном случае) с помощью *минибатчинга* (mini-batching'a) переносятся на задачи стохастической оптимизации[33] [24, 40, 52, 61, 70, 71, 119, 138, 151, 149]. Заме-

[33] Можно обойтись и без минибатчинга, также можно рассмотреть и седловые задачи, см., например, [15, 59, 119, 138, 148, 182, 223]. Отметим, что в работе [182]

тим, что конструкция минибатчинга позволяет переносить оптимальные методы на задачи стохастической оптимизации с сохранением свойства оптимальности и получать оптимальные методы 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 выдает его несмещенную оценку (*стохастический градиент*) $\nabla_x f(x,\xi)$ с конечной дисперсией D:

$$E_\xi\left[\nabla_x f(x,\xi)\right] \equiv \nabla f(x),\ E_\xi\left[\left\|\nabla_x f(x,\xi) - \nabla f(x)\right\|_2^2\right] \le D.$$

Конструкция *минибатчинга* заключается в подстановке в метод вместо неизвестного градиента $\nabla f(x)$ его вычисляемой оценки

$$\overset{r}{\nabla}_x f\left(x, \left\{\xi^l\right\}_{l=1}^r\right) = \frac{1}{r}\sum_{l=1}^r \nabla_x f\left(x,\xi^l\right),$$

где $\left\{\xi^l\right\}_{l=1}^r$ – независимые одинаково распределенные (так же, как ξ) случайные величины, и правильного выбора параметра r. Выбрать этот параметр помогают следующие два неравенства (здесь $L = L_1$ в обозначениях (2.4)):

$$E_{\left\{\xi^l\right\}_{l=1}^r}\left[\left\|\overset{r}{\nabla}_x f\left(x, \left\{\xi^l\right\}_{l=1}^r\right) - \nabla f(x)\right\|_2^2\right] \le \frac{D}{r},$$

$$\left\langle \overset{r}{\nabla}_x f\left(x, \left\{\xi^l\right\}_{l=1}^r\right) - \nabla f(x), \mathrm{v}\right\rangle \le \underbrace{\frac{1}{2L}\left\|\overset{r}{\nabla}_x f\left(x, \left\{\xi^l\right\}_{l=1}^r\right) - \nabla f(x)\right\|_2^2}_{\delta} + \frac{L}{2}\|\mathrm{v}\|_2^2,$$

и результаты о сходимости исследуемого метода при наличии неточного оракула (подобно § 2). Последнее неравенство можно переписать более удобным образом (см. неравенство (2.3)):

$$f\left(x^{k+1}\right) \le f\left(x^k\right) + \left\langle \overset{r}{\nabla}_x f\left(x^k, \left\{\xi^l\right\}_{l=1}^r\right), x^{k+1} - x^k\right\rangle + \frac{2L}{2}\left\|x^{k+1} - x^k\right\|_2^2 + \delta^{k+1}.$$

Используя это неравенство в цепочке рассуждений (2.10) – (2.12), придем к следующему аналогу неравенства (2.12):

описывается метод, который может работать и в условиях отсутствия точных знаний о параметре D.

$$h\left\langle \overset{r}{\nabla}_x f\left(x^k, \left\{\xi^l\right\}_{l=1}^r\right), x^k - x\right\rangle \le h\cdot\left(f\left(x^k\right) - f\left(x^{k+1}\right) + \delta^{k+1}\right) +$$

$$+\frac{1}{2}\left\|x - x^k\right\|_2^2 - \frac{1}{2}\left\|x - x^{k+1}\right\|_2^2,$$

где $h = 1/(2L)$. Беря от обеих частей этого неравенства условное математическое ожидание $E_{x^{k+1}}\left[\ \cdot\ \middle| x^1, ..., x^k\right]$, получим

$$f\left(x^k\right) - f\left(x\right) \le \left\langle \nabla f\left(x^k,\right), x^k - x\right\rangle \le f\left(x^k\right) - E_{x^{k+1}}\left[f\left(x^{k+1}\right)\middle| x^1, ..., x^k\right] +$$

$$+E_{x^{k+1}}\left[\delta^{k+1}\middle| x^1, ..., x^k\right] + L\left\|x - x^k\right\|_2^2 - E_{x^{k+1}}\left[L\left\|x - x^{k+1}\right\|_2^2\middle| x^1, ..., x^k\right].$$

Суммируя выписанные неравенства и беря полное математическое ожидание, можно получить при $x = x_*$ аналог неравенства (2.22) с $L := 2L$. Исходя из (2.22), будем выбирать r следующим образом:

$$\frac{D}{2Lr} \simeq E\left[\delta\right] = \frac{\varepsilon}{2} \ \Rightarrow\ r \simeq \max\left\{\frac{D}{L\varepsilon}, 1\right\}.$$

Поскольку всего итераций (см. § 2 и теорему 3.1)

$$\mathrm{O}\left(\frac{LR^2}{\varepsilon}\right),$$

то общее число обращений к оракулу за стохастическим градиентом $\nabla_x f\left(x, \xi\right)$ при не малых значениях D будет

$$N\left(\varepsilon\right) = \mathrm{O}\left(\frac{DR^2}{\varepsilon^2}\right).$$

Эта же оценка получается и для ускоренных методов. Данная оценка является неулучшаемой оценкой для класса задач выпуклой стохастической оптимизации [52, 80].

Для μ-сильно выпуклой в 2-норме функции $f\left(x\right)$ приведенную оценку можно улучшить с помощью рестартов (см. указание к упражнению 2.3 и конец § 5):

$$N\left(\varepsilon\right) = \mathrm{O}\left(\min\left\{\frac{DR^2}{\varepsilon^2}, \frac{D}{\mu\varepsilon}\right\}\right).$$

Данная оценка является неулучшаемой оценкой для класса задач сильно выпуклой стохастической оптимизации [52, 80]. Заметим, что не сильно

выпуклую часть оценки можно получить из сильно выпуклой с помощью регуляризации $\mu \simeq \varepsilon / R^2$ (см. замечание 4.1).

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)

$$N(\varepsilon) = \mathrm{O}\left(\min\left\{ \frac{\left(L_0^2 + D\right) R^2}{\varepsilon^2}, \frac{L_0^2 + D}{\mu \varepsilon} \right\} \right).$$

Замечание 1. Во многих задачах (в частности, в задачах анализа данных [31, 265, 283]) 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 (объем выборки) m – очень большое число, то вместо честного и дорогого вычисления градиента вычисляют стохастический градиент, случайно (равновероятно) выбирая $r \ll m$ слагаемых $\left\{ \xi^l \right\}_{l=1}^{r}$ и формируя стохастический градиент (несмещенную оценку градиента) по формуле

$$\nabla f\left(x, \left\{ \xi^l \right\}_{l=1}^{r}\right) = \frac{1}{r} \sum_{l=1}^{r} \nabla f_{\xi^l}(x).$$

Такой подход называют *методом рандомизации суммы*, см., например, [18]. С другой стороны, описанную конструкцию можно понимать и как минибатчинг, если посмотреть на исходную постановку задачи следующим образом:

$$f(x) = E_\xi\left[f(x, \xi) \right] \to \min_{x \in Q}, \quad f(x, \xi) = f_\xi(x), \quad \nabla_x f(x, \xi) = \nabla f_\xi(x),$$

$$P(\xi = l) = \frac{1}{m}, \quad l = 1, \ldots, m.$$

Именно в таком ключе обычно смотрят на минибатчинг в глубоком обучении [31, 260].

Отметим, что минибатчинг хорошо параллелится в отличие от процедур типа стохастического усреднения (stochastic averaging) [155]. В популярной работе [245] обсуждается альтернативная возможность распараллеливания стохастического градиентного спуска, в которой отсутству-

ет синхронизационная накладка и обсуждается возможность одновременной записи в общую память.

Продемонстрируем важную роль стохастической оптимизации (рандомизации) при решении *Big Data* задач следующими двумя примерами.

Вернемся 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: $|A_{ij}| \le M$. Если для решения этой задачи использовать быстрый градиентный метод, то оценка общего времени работы метода, необходимого для достижения точности по функции ε будет

$$\underbrace{\mathrm{O}\left(n^2\right)}_{\substack{\text{сложность}\\ \text{итерации}}} \underbrace{\mathrm{O}\left(\sqrt{L_1 R^2/\varepsilon}\right)}_{\text{число итераций}} = \mathrm{O}\left(n^2\sqrt{(M\ln n)/\varepsilon}\right).$$

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

$$\underbrace{\mathrm{O}(n)}_{\substack{\text{сложность}\\ \text{итерации}}} \underbrace{\mathrm{O}\left(\left(L_0^2 R^2\right)/\varepsilon^2\right)}_{\text{число итераций}} = \mathrm{O}\left(n\cdot\left(M^2\ln n\right)/\varepsilon^2\right).$$

Для задач очень больших размеров, при невысоких требованиях к точности решения, второй (рандомизированный) способ может оказаться предпочтительнее.

Вернемся к задаче минимизации суммы из замечания 1 выше. 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. (2.4)). Тогда для решения исходной задачи минимизации суммы с точностью по функции (в среднем) ε требуется $\mathrm{O}\left(L_0^2 R^2/\varepsilon^2\right)$ 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. (2.4)), для достижения той же точности ε быстрому градиентному методу потребуется $\mathrm{O}\left(m\sqrt{L_1R^2/\varepsilon}\right)$ обращений к оракулу за градиентами $f_l(x)$. Для задач с большим числом слагаемых при невысоких требованиях к точности решения первый (рандомизированный) способ может оказаться предпочтительнее. Отметим, что выписанная оценка $\mathrm{O}\left(m\sqrt{L_1R^2/\varepsilon}\right)$ ввиду наличия специальной структуры у задачи уже не будет оптимальной. Оптимальна следующая оценка [85, 88, 205, 210, 211] (в невыпуклом случае см. [86, 87, 90, 204]): $\tilde{\mathrm{O}}\left(m+\sqrt{mL_1R^2/\varepsilon}\right)$. В случае если дополнительно известно, что функция $f(x)$ – μ-сильно выпуклая в 2-норме, то оптимальной оценкой будет: $\tilde{\mathrm{O}}\left(m+\sqrt{mL_1/\mu}\right)$. Эти оценки достижимы, только если имеется доступ к градиенту каждого слагаемого в отдельности [99], а не только к целому градиенту оптимизируемой функции. Далее в примере будет продемонстрирован способ получения такого типа оценок. Тем не менее даже с учетом этого замечания можно привести конкретные примеры, когда первый способ по-прежнему остается предпочтительнее. ■

Все отмеченное выше (до замечания 1) переносится также на покомпонентные и безградиентные постановки задач [7, 13, 15, 18, 20, 52, 61, 71, 153, 156, 228, 243, 267]. Пусть случайный вектор e, например, равномерно распределен на евклидовой сфере в $\mathbb{R}^n$ радиуса 1, т. е. $\|e\|_2=1$ или равновероятно среди единичных ортов [30, 40, 269] (*покомпонентная рандомизация*). Тогда

$$\frac{f(x+\tau e)-f(x)}{\tau}\simeq\langle\nabla f(x),e\rangle,\ E_e\underbrace{\left[n\langle\nabla f(x),e\rangle e\right]}_{\substack{\text{то, что подставляется в}\\ \text{метод вместо градиента}}}=\nabla f(x),$$

$$\langle\nabla f(x),\langle\nabla f(x),e\rangle e\rangle=\left\|\langle\nabla f(x),e\rangle e\right\|_2^2,\ E_e\left[\left\|n\langle\nabla f(x),e\rangle e\right\|_2^2\right]=n\left\|\nabla f(x)\right\|_2^2.$$

Выписанные соотношения вкупе с основным неравенством (3.1) позволяют получить, 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 [52, 81]. Впрочем, при дополнительных предположениях улучшения возможны [7, 13, 20, 153, 244]. 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, вычисляющей значение функции (так бывает далеко не во всех приложениях), то, как правило, лучше попробовать использовать *автоматическое дифференцирование* [246, гл. 8], чем просто аппроксимировать градиент (и тем более старшие производные) конечными разностями [34].

◊ Автоматическое дифференцирование (automatic differentiation) – способ по программе, вычисляющей значение функции (дереву вычислений), построить программу, вычисляющую градиент функции и работающую не дольше, чем в 4 раза, по сравнению с исходной. Однако такой способ требует в общем случае большей памяти – необходимо в памяти хранить всю историю (дерево) вычисления функции. Изначально такого рода результаты были получены (Баур–Штрассен) для полиномов, см. [64] и цитированную там литературу. Впоследствии, в начале 80-х годов XX века, две группы в ЦЭМИ РАН [44] и ВЦ РАН, см. [39] и цитированную там литературу для более полного и точного исторического обзора, смогли получить описанный выше результат в наибольшей общности. Подробный обзор имеется также в работе [106]. В работе [231] приведен 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). Заметим, что аналогом автоматического дифференцирования для негладких выпуклых функций является лексикографическое дифференцирование, предложенное в конце 80-х годов XX века Ю. Е. Нестеровым [232]. 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 счет использования персептронов / функций активаций ReLu) используется процедура автоматического дифференцирования, обоснованно работающая, только для гладких функций. ◊

Далее с помощью техники рестартов, следуя [20], объясняется более точно, откуда в оценке скорости сходимости появляется множитель $\sim n$. Ключевое наблюдение базируется на формуле (1.8):

$$D = E_e\left[\left\|n\langle\nabla f(x), e\rangle e - \nabla f(x)\right\|_2^2\right] \le E_e\left[\left\|n\langle\nabla f(x), e\rangle e\right\|_2^2\right] =$$

$$= n\left\|\nabla f(x)\right\|_2^2 \le 2Ln\cdot\left(f(x) - f(x_*)\right)$$

и приведенной выше оценке скорости сходимости градиентного спуска с минибатчингом для задач стохастической оптимизации (здесь, как

и раньше, работаем с точностью до поправки на вероятности больших уклонений или оговорки о том, что сходимость понимается в среднем, см., например, [70, 174]):

$$f\left(x^N\right)-f\left(x_*\right)=\mathrm{O}\left(\sqrt{\frac{DR^2}{N}}\right)\le\mathrm{O}\left(\sqrt{\frac{2Ln\cdot\left(f\left(x^0\right)-f\left(x_*\right)\right)R^2}{N}}\right).$$

Рестартуя метод, каждый раз, когда происходит гарантированное (выписанной формулой) уполовинивание невязки по функции, получим следующую формулу для общего числа обращений к оракулу за $\left\langle\nabla f\left(x\right),e\right\rangle$ или $\left\{f\left(x\right),f\left(x+\tau e\right)\right\}$ (см. также указания к упражнениям 1.3, 2.3):

$$N\left(\varepsilon\right)=\mathrm{O}\left(n\frac{8LR^2}{\Delta f}\right)+\mathrm{O}\left(n\frac{8LR^2}{\Delta f/2}\right)+\mathrm{O}\left(n\frac{8LR^2}{\Delta f/4}\right)+\ldots$$
$$\ldots+\mathrm{O}\left(n\frac{8LR^2}{\varepsilon}\right)=n\cdot\mathrm{O}\left(\frac{LR^2}{\varepsilon}\right),$$

где $\Delta f=f\left(x^0\right)-f\left(x_*\right)$. К сожалению, по той же причине – грубость формулы (1.8), по 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 [20, 156, 228, 243, 244].

Приведенные выше рассуждения можно повторить в случае μ-сильно выпуклой в 2-норме функции $f\left(x\right)$. Причем все получится еще проще. Можно не повторять рассуждения, а улучшить приведенную оценку с помощью рестартов (см. указание к упражнению 2.3 и конец § 5):

$$N\left(\varepsilon\right)=n\cdot\mathrm{O}\left(\frac{L}{\mu}\left\lceil\ln\left(\frac{\mu R^2}{\varepsilon}\right)\right\rceil\right).$$

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 обнару-

жено [146, 210, 211] (см. также § 3), как приведенные 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 *Катализ*т:

$$N(\varepsilon) = n \cdot \mathrm{O}\left(\min\left\{ \sqrt{\frac{LR^2}{\varepsilon}}, \sqrt{\frac{L}{\mu}} \left\lceil \ln\left(\frac{\mu R^2}{\varepsilon} \right) \right\rceil \right\} \right).$$

В основе подхода Каталист лежит ускоренный прокс-метод, полученный в концепции модели функции [72]. Неускоренный вариант разбирался в § 3. Основным «структурным блоком» по-прежнему является вспомогательная задача (3.21),

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

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 $\tilde{\mathrm{O}}\left(n \cdot \left(L + \tilde{L} \right) \middle/ \left(\mu + \tilde{L} \right) \right)$. С другой стороны, число внешних итераций ускоренного прокс-метода [72, 210, 211] с точностью до логарифмического множителя равно $\tilde{\mathrm{O}}\left(\sqrt{\tilde{L}/\mu} \right)$. Оценка общего числа обращений к оракулу будет наилучшей $\tilde{\mathrm{O}}\left(n\sqrt{L/\mu} \right)$ при выборе $\tilde{L} \simeq L$, что соответствует в сильно выпуклом случае ранее приведенному результату.

◊ Заметим, из того, что описано в § 3 с помощью рестартов (см. конец § 5), можно получить только оценку $\tilde{\mathrm{O}}\left(\tilde{L}/\mu \right)$ на число внешних итераций. Для общего ускорения необходимо, чтобы внешний метод был ускоренным, см. упражнение 3.7! ◊

Разобранный пример наглядно демонстрирует общую идею подхода: за счет выбора параметра регуляризации $\tilde{L}$ добиваться близкой к единице обусловленности вспомогательной задачи, тогда неоптимальность используемого для ее решения подхода (напомним, что как раз для внутренней задачи используется неускоренный метод, который хотим ускорить) становится несущественной, и за счет ускоренности внешнего метода происходит общее ускорение рассматриваемой процедуры. Примечательно, что внешний ускоренный прокс-метод остается одним и тем же в дан-

ном подходе, 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 оптимальным с точностью до логарифмических (по желаемой точности) множителей.

Каталист хотя и является универсальным способом ускорения всевозможных неускоренных методов, тем не менее на практике предпочитают использовать прямые ускоренные рандомизированные процедуры, см., например, [85, 205, 206] и замечание 2 ниже. Особенно активно в этом направлении работают З. Аллен-Зу [84] и Дж. Лан [198].

Важно также отметить, что даже в неускоренном случае приведенные выше рассуждения при покомпонентной рандомизации оказываются достаточно грубыми, поскольку не учитывают, что константа L теперь можно считать не по худшему направлению, а «средней» по всем направлениям, что может быть в $\sim\sqrt{n}$ раз меньше [20, 244].

Замечание 2. Предположим, что для всех $x\in\mathbb{R}^n$ и $h\in\mathbb{R}$:

$$\left|\partial f\left(x+he_i\right)/\partial x_i-\partial f\left(x\right)/\partial x_i\right|\le L_i h.$$

Пусть $\beta\in[0,1]$. Введем

$$\left\|x\right\|^2=\sum_{i=1}^n L_i^{1-2\beta}x_i^2,\ \breve{R}^2=\frac{1}{2}\left\|x_*-x^0\right\|^2,\ \left\|\nabla f\left(x\right)\right\|_*^2=\sum_{i=1}^n L_i^{2\beta-1}\cdot\left(\frac{\partial f\left(x\right)}{\partial x_i}\right)^2.$$

Метод линейного каплинга (МЛК)	Покомпонентный вариант МЛК
См. указание к упражнению 1.3 $x^{k+1}=\tau z^k+(1-\tau)y^k,$ $y^{k+1}=x^{k+1}-\frac{1}{L}\nabla f\left(x^{k+1}\right),$ $z^{k+1}=z^k-h\nabla f\left(x^{k+1}\right).$	$x^{k+1}=\tau z^k+(1-\tau)y^k,$ Случайно и независимо разыгрываем $i_{k+1}\in[1,\dots,n]$ по правилу: $P\left(i_{k+1}=i\right)=p_i\overset{def}{=}\frac{L_i^\beta}{\sum_{j=1}^n L_j^\beta},\ i=1,\dots,n,$ $y_{i_{k+1}}^{k+1}=x_{i_{k+1}}^{k+1}-\frac{1}{L_{i_{k+1}}}\frac{\partial f\left(x^{k+1}\right)}{\partial x_{i_{k+1}}^{k+1}},$ $z_{i_{k+1}}^{k+1}=z_{i_{k+1}}^{k}-\frac{h}{p_{i_{k+1}}}\frac{\partial f\left(x^{k+1}\right)}{\partial x_{i_{k+1}}^{k+1}}.$

Для МЛК согласно указанию к упражнению 1.3 имеем

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

т. е.

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

Для ПМЛК аналогом приведенных неравенств будут

$$\frac{1}{p_{i_{k+1}}}\left\langle \left\langle \nabla f\left(x^{k+1}\right), e_{i_{k+1}} \right\rangle e_{i_{k+1}}, z^k - x_* \right\rangle \le$$

$$\le \frac{1}{2h}\left\|z^k - x_*\right\|^2 - \frac{1}{2h}\left\|z^{k+1} - x_*\right\|^2 + \frac{h\left\|\left\langle \nabla f\left(x^{k+1}\right), e_{i_{k+1}} \right\rangle e_{i_{k+1}}\right\|_*^2}{2p_{i_{k+1}}^2},$$

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

$$-E_{i_{k+1}}\left[\frac{1}{2h}\left\|z^{k+1} - x_*\right\|^2 \middle| i_0,...,i_k\right] + \breve{L}h\cdot\left(f\left(x^{k+1}\right) - E_{i_{k+1}}\left[f\left(y^{k+1}\right)\middle| i_0,...,i_k\right]\right),$$

где $\breve{L} = \left(\sum_{j=1}^{n} L_j^{\beta}\right)^2$. Второе неравенство получается из первого взятием условного математического ожидания от обеих частей $E_{i_{k+1}}\left[\,\cdot\,\middle| i_0,...,i_k\right]$. Поскольку согласно указанию к упражнению 1.3 оценка числа итераций (вычислений градиента $f(x)$), необходимых для достижения по функции точности ε, имеет вид: $\mathrm{O}\left(\sqrt{LR^2/\varepsilon}\right)$, то 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 $\mathrm{O}\left(\sqrt{\breve{L}\breve{R}^2/\varepsilon}\right)$. Так оно в действительности и оказывается [20, 93, 228, 244]. Аналогичные рассуждения можно было провести, взяв за основу метод подобных треугольников вместо МЛК [156].

Метод МЛК и ПМЛК можно осуществлять (с такими же оценками скорости сходимости) и без рестартов [20, 92, 93]. Для этого нужно выбирать:

$$\tau_k = 2/(k+2) \text{ и } h_k = (k+2)/(2L) \text{ (МЛК)},$$

$$\tau_k = 2/(k+2) \text{ и } h_k = (k+2)/(2\breve{L}) \text{ (ПМЛК)}.$$

Описанный покомпонентный метод имеет естественное блочно-покомпонентное обобщение [228]. При этом допускается, что рассматриваемую задачу оптимизации необходимо решать на множестве простой структуры, имеющей вид прямого произведения множеств, отвечающих различным блокам [20, 156]. ■

Поясним написанное простым примером, в котором сравним время работы быстрого градиентного метода, например, из указания к упражнению 1.3 – МЛК, и его покомпонентного варианта – ПМЛК с $\beta = 1/2$ [20]. То же самое можно было продемонстрировать и для неускоренных методов. Итак, рассматривается задача квадратичной выпуклой оптимизации (1.30) в условиях работы [244]:

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

где $A = \|A_{ij}\|_{i,j=1}^{n} \succ 0$ и $1 \le A_{ij} \le 2$ при $i, j = 1, ..., n$. Из последнего условия имеем

$$L = \lambda_{\max}(A) \ge \lambda_{\max}(1_n 1_n^T) = n,$$

поэтому оценка общего времени работы МЛК (оптимального, с точностью до числового множителя, метода для данного класса задач) будет

$$\underbrace{\mathrm{O}(n^2)}_{\substack{\text{стоимость}\\ \text{итерации}}} \cdot \underbrace{\mathrm{O}\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)}_{\substack{\text{число}\\ \text{итерации}}} \sim \mathrm{O}\left(\frac{n^{5/2}}{\varepsilon^{1/2}}\right).$$

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 $\sqrt{L} \ge \sqrt{n}$ можно ставить

$$\sqrt{\overline{L}} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{L_i} \le \sqrt{2},$$

т. е. число итераций согласно замечанию 1 (см. также [20, 244]) будет

$$\mathrm{O}\left(n\sqrt{\frac{\bar{L}R^2}{\varepsilon}}\right) \sim \mathrm{O}\left(\frac{n}{\varepsilon^{1/2}}\right).$$

При этом стоимость итерации теперь будет $\mathrm{O}(n)$. Действительно, если $\tilde{x} = x + he_i$, где e_i – i-орт и уже посчитано Ax, то $A\alpha\tilde{x} = \alpha Ax + \alpha hA^{\langle i\rangle}$, где α – заданное число, может быть посчитано за $\mathrm{O}(n)$. Отсюда с помощью 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:

$$\underbrace{\mathrm{O}(n)}_{\substack{\text{стоимость}\\ \text{итерации}}} \cdot \underbrace{\mathrm{O}\left(n\sqrt{\frac{\bar{L}R^2}{\varepsilon}}\right)}_{\substack{\text{число}\\ \text{итерации}}} \sim \mathrm{O}\left(\frac{n^2}{\varepsilon^{1/2}}\right) \ll \mathrm{O}\left(\frac{n^{5/2}}{\varepsilon^{1/2}}\right).$$

Кажется, что получается какое-то противоречие с оптимальностью МЛК для класса гладких выпуклых задач и противоречие с рядом тезисов, которые ранее приводились в пособии. На самом деле никаких противоречий нет [52]:

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 трудоемкости). По времени работы не существует теории нижних оценок и вряд ли в ближайшее время можно ожидать ее появления.
4) Наконец, ранее отмечалось, что если есть возможность считать градиент, например, с помощью автоматического дифференцирования, то стоит именно градиент и использовать в методе, а не восстанавливать компоненты градиента по посчитанным значениям функции. Однако в

последнем примере используется не процедура расчета значений функции, а процедура их пересчета. Действительно, вычислив самый первый раз $f(x)$ за $O(n^2)$, 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 приближенно сделано за два пересчета значения функции, т. е. за время $O(n)$.

Может показаться, что выше слишком много внимания было уделено, на первый взгляд, довольно незначительной оговорке о возможности в специальных (ассиметричных) случаях ускорять время работы методов в $\sim\sqrt{n}$ раз за счет 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 счет самого факта рандомизации, а не за счет ее специфики [20], в общем случае следует признать удачными в оптимизационной практике приемы типа диагонального шкалирования (предобуславливание), пришедшие из вычислительной линейной алгебры [28, 69, 167, 278], 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 [65, 77]. Однако наиболее популярными сейчас являются *квазиньютоновские методы* [246] (см. также замечание 3) и методы типа *AdaGrad* [31, 148, 260].

Однако главная причина такого внимания к покомпонентным методам описана ниже. Исследования последних лет показывают (см., например, [20] 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 [265]. В частности, различные варианты метода рандомизации суммы (см. замечание 1) можно понимать, как варианты покомпонентных рандомизаций для двойственной задачи.

Пример (типичная задача анализа данных [20, 88, 115, 265, 266]). Рассмотрим задачу выпуклой оптимизации:

$$\sum_{k=1}^{m} f_k\left(A_k^T x\right) + g(x) \to \min_{x \in Q},$$

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), Q – множество простой структуры. Предполагаем, что трудоемкость вычисления $f_k'(z_k)$ равна $\mathrm{O}(1)$ и что для всех $k = 1, ..., m$ и всех допустимых w, v:

$$\left| f_k'(w) - f_k'(v) \right| \le L|w - v|.$$

Функция $g(x)$ предполагается сильно выпуклой в p-норме с константой μ_p. Вводя матрицу $A = \left[A_1, ..., A_m\right]^T$ и вспомогательный вектор $z = Ax$, можно переписать эту задачу в «раздутом» пространстве $\tilde{y} = (x, z)$, как задачу типа проектирования на аффинное многообразие [2, 97, 147], см. также (4.7). Эта задача проектирования решается путем перехода к двойственной задаче. Опишем далее довольно общий способ построения двойственной задачи:

$$\min_{x \in Q}\left\{\sum_{k=1}^{m} f_k\left(A_k^T x\right) + g(x)\right\} = \min_{\substack{x \in Q \\ z = Ax}}\left\{\sum_{k=1}^{m} f_k(z_k) + g(x)\right\} =$$

$$= \min_{\substack{x \in Q \\ z = Ax,\, z'}} \max_{y}\left\{\langle z - z', y\rangle + \sum_{k=1}^{m} f_k(z_k') + g(x)\right\} =$$

$$= \max_{y \in \mathbb{R}^m}\left\{-\max_{\substack{x \in Q \\ z = Ax}}\left\{\langle -z, y\rangle - g(x)\right\} - \max_{z'}\left\{\langle z', y\rangle - \sum_{k=1}^{m} f_k(z_k')\right\}\right\} =$$

$$= \max_{y \in \mathbb{R}^m}\left\{-\max_{x \in Q}\left(\langle -A^T y, x\rangle - g(x)\right) - \sum_{k=1}^{m} \max_{z_k'}\left(z_k' y_k - f_k(z_k')\right)\right\} =$$

$$= \max_{y \in \mathbb{R}^m}\left\{-g^*\left(-A^T y\right) - \sum_{k=1}^{m} f_k^*(y_k)\right\} = -\min_{y \in \mathbb{R}^m}\left\{g^*\left(-A^T y\right) + \sum_{k=1}^{m} f_k^*(y_k)\right\}.$$

В описанную схему погружаются, например, следующие задачи [97]:

1) $\dfrac{L}{2}\left\|Ax-b\right\|_2^2+\underbrace{\dfrac{\mu}{2}\left\|x-x_g\right\|_2^2}_{g(x)}\to\min\limits_{x\in\mathbb{R}^n}$, (Ridge regression)

2) $\dfrac{L}{2}\left\|Ax-b\right\|_2^2+\underbrace{\mu\sum\limits_{k=1}^{n}x_k\ln x_k}_{g(x)}\to\min\limits_{x\in S_n(1)}$. (Minimal mutual information model)

Константа Липшица производной $f_k(z)=\dfrac{L}{2}(z_k-b_k)^2$ равна L, константы сильной выпуклости $g(x)$ (считаются в разных нормах: 1) в 2-норме, 2) в 1-норме) также одинаковы в обоих случаях и равны μ. Для приведенных выше задач получим следующие двойственные задачи:

1) $\dfrac{1}{2\mu}\left(\left\|x_g-A^Ty\right\|_2^2-\left\|x_g\right\|_2^2\right)+\dfrac{1}{2L}\left(\left\|y+b\right\|_2^2-\left\|b\right\|_2^2\right)\to\min\limits_{y\in\mathbb{R}^m}$,

2) $\dfrac{1}{\mu}\ln\left(\sum\limits_{i=1}^{n}\exp\left(\dfrac{\left[-A^Ty\right]_i}{\mu}\right)\right)+\dfrac{1}{2L}\left(\left\|y+b\right\|_2^2-\left\|b\right\|_2^2\right)\to\min\limits_{y\in\mathbb{R}^m}$.

В общем случае можно утверждать, что $\sum\limits_{k=1}^{m}f_k^*(y_k)$ (композитный член в двойственной задаче, см. пример 3.1) является сильно выпуклым в 2-норме с константой сильной выпуклости равной L^{-1}. Для $g^*\left(-A^Ty\right)$ можно оценить константу Липшица градиента в 2-норме (см. конец § 2, начало § 4, а также указание к упражнению 4.8):

$$\frac{1}{\mu}\max_{\|y\|_2\le1,\,\|x\|_p\le1}\left\langle A^Ty,x\right\rangle^2=\frac{1}{\mu}\max_{\|x\|_p\le1}\left\|Ax\right\|_2^2=\frac{1}{\mu}\begin{cases}1)\ \lambda_{\max}\left(A^TA\right)\\ 2)\ \max\limits_{j=1,\dots,n}\left\|A^j\right\|_2^2\end{cases}$$

и получить следующую оценку сверху на константы Липшица всех частных производных $g^*\left(-A^Ty\right)$ [20]:

$$\frac{1}{\mu}\max_{\|y\|_1\le1,\,\|x\|_p\le1}\left\langle A^Ty,x\right\rangle^2=\frac{1}{\mu}\max_{\|x\|_p\le1}\left\|Ax\right\|_\infty^2=\frac{1}{\mu}\begin{cases}1)\ \max\limits_{i=1,..,m}\left\|A_i\right\|_2^2\\ 2)\ \max\limits_{\substack{i=1,\dots,m\\ j=1,\dots,n}}\left|A_{ij}\right|^2\end{cases}.$$

Для ПМЛК с $\beta=0$ из замечания 2, примененного к двойственной задаче (в случае, когда матрица A плотно заполненная), имеем следующие оценки общего времени работы (трудоемкости):

1) $T_1 = \tilde{\mathrm{O}}\left(n \cdot m \sqrt{\frac{L \max\limits_{i=1,..,m} \|A_i\|_2^2}{\mu}} \right)$,

2) $T_2 = \tilde{\mathrm{O}}\left(n \cdot m \sqrt{\frac{L \max\limits_{i,j} |A_{ij}|^2}{\mu}} \right)$.

Если теперь посмотреть на исходную прямую задачу (с $L := L/m$):

$$\frac{1}{m} \sum_{k=1}^{m} f_k\left(A_k^T x\right) + g(x) \to \min_{x \in Q},$$

и оценить трудоемкость оптимальных методов согласно оценкам, приведенным в конце замечания 1, то получим уже анонсированное соответствие оптимальных оценок с полученными только что оценками трудоемкости ПМЛК с $\beta = 0$ (при $L := L/m$). Действительно, с учетом того, что константы Липшица градиентов $f_k\left(A_k^T x\right)$, посчитанные в соответствующих нормах (соответствующих норме, в которой сильно выпукл композит прямой задачи), равномерно (по $k = 1, ..., m$) оцениваются следующим образом:

1) $L \max\limits_{i=1,..,m} \|A_i\|_2^2$,

2) $L \max\limits_{i,j} |A_{ij}|^2$,

а сложность вычисления $\nabla f_k\left(A_k^T x\right)$ равна $\mathrm{O}(n)$, то в типичном случае:

1) $\tilde{T}_1 = \tilde{\mathrm{O}}\left(n \cdot \left(m + \sqrt{m \frac{L \max\limits_{i=1,..,m} \|A_i\|_2^2}{\mu}} \right) \right) = \tilde{\mathrm{O}}\left(n \cdot m \sqrt{\frac{(L/m) \max\limits_{i=1,..,m} \|A_i\|_2^2}{\mu}} \right)$,

2) $\tilde{T}_2 = \tilde{\mathrm{O}}\left(n \cdot \left(m + \sqrt{m \frac{L \max\limits_{i,j} |A_{ij}|^2}{\mu}} \right) \right) = \tilde{\mathrm{O}}\left(n \cdot m \sqrt{\frac{(L/m) \max\limits_{i,j} |A_{ij}|^2}{\mu}} \right)$.

Таким образом, имеет место полное соответствие (с точностью до опущенных в рассуждениях логарифмических множителей). Интересно заметить, что для первой задачи здесь можно сполна использовать разреженность матрицы A [20]. Более того, эту задачу (также с полным учетом разреженности) можно решать и прямым ПМЛК с $\beta = 0$. Соответствую-

щие оценки согласно замечанию 2 имеют вид (снова возвращаемся к исходному пониманию параметра L):

$$T_1^{\text{прям}} = \tilde{\mathrm{O}}\left(s \cdot n \sqrt{\frac{L \max\limits_{j=1,\ldots,n} \left\| A^j \right\|_2^2}{\mu}} \right),$$

$$T_1^{\text{двойств}} = \tilde{\mathrm{O}}\left(\tilde{s} \cdot m \sqrt{\frac{L \max\limits_{i=1,\ldots,m} \left\| A_i \right\|_2^2}{\mu}} \right) = \tilde{\mathrm{O}}\left(sn \sqrt{\frac{L \max\limits_{i=1,\ldots,m} \left\| A_i \right\|_2^2}{\mu}} \right),$$

где s – среднее число ненулевых элементов в столбцах матрицы A, а $\tilde{s}$ – в строках.

В действительности, обе эти оценки оказываются завышенными.[34] Более аккуратные рассуждения позволяют заменить максимум в этих оценках на некоторые средние. Скажем, в транспортных приложениях [15, гл. 1], [97] матрица A не просто разреженная, но еще и битовая (состоит из нулей и единиц). В таком случае приведенные оценки переписываются следующим образом:

$$T_1^{\text{прям}} = \tilde{\mathrm{O}}\left(sn \sqrt{\frac{Ls}{\mu}} \right), \; T_1^{\text{двойств}} = \tilde{\mathrm{O}}\left(sn \sqrt{\frac{L\tilde{s}}{\mu}} \right).$$

Отсюда можно сделать довольно неожиданный вывод [97]: при $m \ll n$ стоит 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. ■

Одним из наиболее активных специалистов по рандомизированным методам выпуклой оптимизации сейчас является П. Рихтарик [255].

◊ Отметим, что описанное выше направление в его современном варианте появилось в связи с двумя препринтами Ю. Е. Нестерова, подготовленными в 2010–2011 гг. Результаты, приведенные в этих препринтах, впоследствии вошли в две статьи [228, 243]. 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 распределенным вариантам) всевозможные приложе-

[34] 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.

ния, особенно в задачах огромных размеров (*Big Data*), приходящих из анализа данных. ◊

В описанной выше общности 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 и задач выпуклой оптимизации с оракулом, выдающем модель функции и их (блочно) покомпонентные варианты) позволяют строить методы, работающие по наилучшим известным сейчас теоретическим оценкам общего времени работы (трудоемкости) практически для всех известных нам классов задач (структурной) выпуклой оптимизации больших размеров.

Большое число исследований во всем мире сейчас сосредоточено на изучении (переборе) конкретных способов сочетания описанных выше приемов 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, «ускоряющие» почти все оценки, приведенные 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 приемы типа *альтернативных направлений*, *метода штрафных функций*, *метода модифицированной функции Лагранжа*, *ADMM* и др. [11, 36, 41, 61, 67, 107, 202, 203, 246, 249, 284] 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 счет проксимальной природы ряда описанных процедур, например, метода модифицированной функции Лагранжа [277], см. также § 3, упражнение 4.3 и замечание 4.3. Но сполна этим можно воспользоваться, как правило, только в случае организации распределенных вычислений, см., например, ADMM [115, 249, 284].

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 [39, 106, 246] и возможность быстрого пересчета значений функции, (компонент) градиента [15, 237]. В этой связи достаточно привести несколько примеров [15, п. 1.5.2, 4, 5.1], [119, п. 3.3], [133, 186, 237, 242], см. также упражнение 1.6. Все эти при-

меры ярко демонстрируют, что на практике за счет дешевых 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 зашумлен [12, 38, 39, 161, 250]. Наконец, почти ничего не было сказано о многих других методах и подходах, которые часто используются при решении практических задач умеренных размеров [160]: в частности, квазиньютоновских методах [246] и методах второго порядка (ньютоновские методы), интерес к которым в последние годы резко возрос [27, 39, 53, 54, 82, 91, 101, 102, 121, 132, 164, 171, 172, 217, 227, 231, 239, 283].

Замечание 3 (квазиньютоновские методы [246, 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 все-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* [61, п. 2 § 3, гл. 3] (варианта метод сопряженных градиентов). Наиболее интересными в практическом плане являются способы, которые требуют не более чем квадратичной (по размерности пространства) трудоемкости и памяти. Среди таких способов наиболее удачно себя зарекомендовал способ, приводящий в итоге к методу *BFGS*:

$$h_k = \arg\min_{h \in \mathbb{R}} f\left(x^k - hH_k \nabla f\left(x^k\right)\right),$$

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

$$H_{k+1} = H_k + \frac{H_k\gamma_k\delta_k^T + \delta_k\gamma_k^T H_k}{\left\langle H_k\gamma_k, \gamma_k\right\rangle} - \beta_k \frac{H_k\gamma_k\gamma_k^T H_k}{\left\langle H_k\gamma_k, \gamma_k\right\rangle},$$

где

$$\beta_k = 1 + \frac{\left\langle \gamma_k, \delta_k\right\rangle}{\left\langle H_k\gamma_k, \gamma_k\right\rangle},\ \gamma_k = \nabla f\left(x^{k+1}\right) - \nabla f\left(x^k\right),\ \delta_k = x^{k+1} - x^k,\ H_0 = I.$$

В отличие от сопряженных 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 [61, п. 4 § 4, гл. 5] (метод эллипсоидов, методы Шора). Рисунок 4 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 растяжение/сжатие и позволяют адаптивно улучшать обусловленность задачи, правильно аккумулируя собранную информацию в H_k.

В теоретическом плане по квазиньютоновским методам на данный момент известно не так уж и много (см., например, [54, п. 1.3.1]): 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.

Основным ограничением по использованию квазиньютоновских методов является необходимость в хранении и обновлении плотной квадратной матрицы H_k, что требует (в отличие от того, что имеет место для методов типа сопряженных градиентов) квадратичной памяти и квадратичного времени независимо от разреженности задачи. Это обстоятельство существенно ограничивает возможности по использованию 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* [246, 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 растет. Поэтому обычно последовательности векторов $\{\gamma_l\}_{l=0}^k$ и $\{\delta_l\}_{l=0}^k$ хра-

нят только с q последних итераций (q – глубина памяти), и при этом полагают $H_{k-q} = I$. На практике q часто выбирают совсем небольшим: $q \simeq 3-5$.

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 [246, гл. 7] и всевозможных рандомизированных вариантов квазиньютоновских методов [112, 169, 260]. ■

В заключение все же приведем сопоставительный анализ методов первого порядка (градиентных методов) и методов более высокого порядка, которые могут использоваться для решения задач выпуклой оптимизации умеренных размеров ($n \le 10^4$), в условиях отсутствия шума на классе достаточно гладких

$$\left\| \nabla^r f(y) - \nabla^r f(x) \right\|_2 \le M_r \left\| y - x \right\|_2, \ x, y \in \mathbb{R}^n, \ M_r \le \infty, \ r = 0,1,2, \ldots,$$

μ-сильно выпуклых в 2-норме задач, где $\mu \ge 0$.

◊ 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

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

$$\left\| \nabla^2 f(y) - \nabla^2 f(x) \right\|_2 = \sup_{\|x_1\|_2 \le 1} \sup_{\|x_2\|_2 \le 1} \left\langle \left(\nabla^2 f(y) - \nabla^2 f(x) \right)[x_1], x_2 \right\rangle =$$

$$= \sup_{\|x_1\|_2 \le 1} \sup_{\|x_2\|_2 \le 1} \left\langle \left(\nabla^2 f(y) - \nabla^2 f(x) \right) x_1, x_2 \right\rangle =$$

$$= \max \left\{ \lambda_{\max} \left(\nabla^2 f(y) - \nabla^2 f(x) \right), \left| \lambda_{\min} \left(\nabla^2 f(y) - \nabla^2 f(x) \right) \right| \right\}.$$

В общем случае см. [102, 231]. Отметим также, что при $r = 0$: $\nabla^0 f(x) = f(x)$, 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, (2.4)), которая только в худшем случае совпадает с введенной здесь [233]. Аналогичное замечание имеет место и по методам 2-го порядка [239] (и, вероятно, более высокого порядка). «Правильный» метод p-го порядка ($p \ge 1$) на первых итерациях «осуществляет» желаемую редукцию (уменьшение) констант гладкости

$\{M_r\}_{r=0}^{p-1}$ за счет попадания в нужную область сходимости метода. Причем часто достаточно одной (первой) итерации [233, 239]. ◊

Для класса методов, у которых на каждой итерации разрешается не более чем $O(1)$ раз обращаться к оракулу (подпрограмме) за $\nabla^r f(x)$, $r \le 1$, оценка числа итераций, необходимых для достижения точности ε (по функции), будет иметь вид

$$O\left(\min\left\{ n\ln\left(\frac{\Delta f}{\varepsilon}\right);\ \frac{M_0^2 R^2}{\varepsilon^2}, \left(\frac{M_1 R^2}{\varepsilon}\right)^{1/2};\ \frac{M_0^2}{\mu\varepsilon}, \left(\frac{M_1}{\mu}\right)^{1/2} \left\lceil \ln\left(\frac{\mu R^2}{\varepsilon}\right)\right\rceil \right\} \right),$$

где, как и раньше, $R = \left\| x^0 - x_* \right\|_2$, $\Delta f = f(x^0) - f(x_*)$. Данная оценка в общем случае не может быть улучшена даже если дополнительно известно, что, $M_2 < \infty$, $M_3 < \infty$, … [52]. При этом данная оценка достигается [52, 54, 119].

◊ Заметим, что если вместо $r \le 1$ имеет место $r = 0$ (класс безградиентных методов), то в приведенной оценке все аргументы минимума следует домножить на размерность пространства n [7, 13, 15, 52, 63, 152, 153, 156, 243]. 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 [52, 119, 209]). ◊

Для класса методов, у которых на каждой итерации разрешается не более чем $O(1)$ раз обращаться к оракулу (подпрограмме) за значениями $\nabla^r f(x)$, $r \le p$, $p \ge 2$, оценка числа итераций, необходимых для достижения точности ε (по функции), будет иметь вид

$$O\left(\min\left\{ n\ln\left(\frac{\Delta f}{\varepsilon}\right);\ \frac{M_0^2 R^2}{\varepsilon^2}, \left(\frac{M_1 R^2}{\varepsilon}\right)^{1/2}, \left(\frac{M_2 R^3}{\varepsilon}\right)^{2/7}, ..., \left(\frac{M_p R^{p+1}}{\varepsilon}\right)^{2/(3p+1)};\right.\right.$$

$$\left. \min\left\{ \left(\frac{M_1}{\mu}\right)^{1/2}, \left(\frac{M_2 R}{\mu}\right)^{2/7}, ..., \left(\frac{M_p R^{p-1}}{\mu}\right)^{2/(3p+1)} \right\} + \right.$$

$$\left.\left. + \min_{r=2,...,p} \log\left\lceil \log\left(\frac{\left(\mu^{r+1}/M_r^2\right)^{1/(r-1)}}{\varepsilon} \right)\right\rceil \right\} \right).$$

Данная оценка в общем случае не может быть улучшена даже если дополнительно известно, что, $M_{p+1} < \infty$, $M_{p+2} < \infty$, … [52, 101]. Полезно отметить, что здесь, также 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 [101, 231]. При этом недавно было подмечено [101, 217], что данная оценка с точностью до числового множителя в выпуклом случае и с точностью до логарифмического множителя $\log\left(M_1 M_2^2 R^2 / \mu^3\right)$ в μ-сильно выпуклом случае достигается при $p = 2$. В работе [231] также было показано, как при $p = 2,3$ получать методы с почти оптимальными оценками скорости сходимости, трудоемкость каждой итерации которых сопоставима (с точностью до логарифмических множителей) с трудоемкостью итерации метода Ньютона. Наиболее важным наблюдением работы [231] является следующее: для выпуклой функции $f(x)$ сумма, вообще говоря, невыпуклого по y многочлена Тейлора:

$$\sum_{r=0}^{p} \frac{1}{r!} \nabla^r f(x) \underbrace{[y-x, \ldots, y-x]}_{r}$$

и

$$\frac{M}{(p+1)!} \|y-x\|_2^{p+1}, \; M \geq pM_p$$

будет выпуклой по y функцией. В первую очередь, именно это наблюдение позволяет осуществлять каждую итерацию метода за разумное время, во вторую очередь, это концепция относительной точности (см. начало § 3).

◊ В работе [83] приводятся отличные от [101] нижние оценки при более общих условиях. Отметим, что по-прежнему не создано общей теории правильным образом укоренных методов высокого порядка. Ускорение, унаследованное от градиентных методов, не дает оптимальных оценок [102, 227, 231]. Имеются разные гипотезы в этом направлении, например, что подобно указанию к упражнению 1.3 и замечаниям 1.5, 1.6,

для $p=2$ можно попробовать искать оптимальные методы в классе трехшаговых процедур или процедур с небольшим числом пересчитываемых последовательностей, использующих информацию: $\nabla f\left(x\right)$, $\nabla^2 f\left(x\right)$. Однако каких-то законченных результатов тут на данный момент не известно.

Важно подчеркнуть, что нет сомнений в правильности (достижимости с точностью до логарифмических множителей) оценки на число итераций, приведенной выше. Здесь хотелось лишь обратить внимание, что на данный момент не удалось в полном объёме (математически строго) обосновать достижимость этой оценки во всех случаях, т. е. при $p>2$.

Отметим также, что приведенная оценка (случай $p=2$), по-видимому, была известна А. С. Немировскому еще в 80-е годы XX века. Однако она не была опубликована, потому что тогда не удалось найти методы, которые бы так работали. В ближайшее время можно ожидать, что в этом направлении будут получены значительные продвижения. ◊

Поясним, как получается сильно выпуклая часть последней оценки из не сильно выпуклой.[35] Для этого рассмотрим *метод Ньютона*[36]:

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

$$=x^k-\left[\nabla^2 f\left(x^k\right)\right]^{-1}\nabla f\left(x^k\right).$$

[35] 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).

[36] В методе Ньютона на каждой итерации предлагается минимизировать параболоид (уже, вообще говоря, не вращения), касающийся (до второго порядка включительно, то есть 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 такого типа методов второго порядка является добавление к параболоиду кубического слагаемого – кубической регуляризации [102, 227, 239]. Полученная в результате вспомогательная задача (минимизации регуляризованного кубическим членом параболоида) по сложности сопоставима с исходной [124, 239], но при этом метод приобретает глобальную сходимость. Отметим, что локальная сходимость остается по-прежнему квадратичной, как у метода Ньютона.

Считая, что $M_2 < \infty$, $\mu > 0$, получим

$$\left\|\nabla f\left(x^{k+1}\right)\right\|_2 = \left\|\nabla f\left(x^{k+1}\right) - \nabla f\left(x^k\right) - \nabla^2 f\left(x^k\right)\left(x^{k+1} - x^k\right)\right\|_2 \le M_2 \left\|x^{k+1} - x^k\right\|_2^2 =$$

$$= M_2 \left\|\left[\nabla^2 f\left(x^k\right)\right]^{-1} \nabla f\left(x^k\right)\right\|_2^2 \le \frac{M_2}{\mu^2}\left\|\nabla f\left(x^k\right)\right\|_2^2.$$

Заметим, что если последовательность положительных чисел $\{c_k\}_{k=0,1,2,\ldots}$ удовлетворяет условию (обеспечивающему *квадратичную скорость сходимости*):

$$c_{k+1} \le \text{const} \cdot \left(c_k\right)^{\gamma},\ \gamma > 1,$$

и c_0 достаточно мало, то после $N = \mathrm{O}\left(\log\left\lceil \log\left(c_0/\varepsilon\right)\right\rceil\right)$ итераций $c_N \le \varepsilon$. Для метода Ньютона $c_k = \left\|\nabla f\left(x^k\right)\right\|_2$, $\gamma = 2$; для класса чебышёвских методов высокого порядка $c_k = \left\|x^k - x_*\right\|_2$, $\gamma = 3, 4, 5, \ldots$ [39, п. 2.9], [43, п. 9.5.10]; для методов Ньютона с кубической регуляризацией $c_k = f\left(x^k\right) - f\left(x_*\right)$, $\gamma = 4/3$, причем в последнем случае сильную выпуклость можно заменить градиентным доминированием [53, гл. 4], [227, 239] (см. также замечание 1.1).

С помощью неравенства (1.15) для метода Ньютона можно оценить окрестность квадратичной скорости сходимости метода:

$$\frac{M_2}{\mu^2}\left\|\nabla f\left(x^k\right)\right\|_2 < 1 \ \Rightarrow\ f\left(x^k\right) - f\left(x_*\right) \le \frac{1}{2\mu}\left\|\nabla f\left(x^k\right)\right\|_2^2 < \frac{\mu^3}{2M_2^2}.$$

Оказавшись в этой окрестности, можно достичь желаемой точности за $\log\left\lceil \log\left(\left(\mu^3/M_2^2\right)/\varepsilon\right)\right\rceil$ итераций. Чтобы оказаться в этой окрестности, можно использовать технику рестартов (см. упражнение 2.3, § 5, а также [53, гл. 4], [101, 227]), примененную к методам, обеспечивающим несильно выпуклую составляющую рассматриваемой оценки. Таким образом и получается вторая заключительная (сильно выпуклая) часть рассматриваемой оценки. Точнее говоря, получается вот такая оценка [101]:

$$\mathrm{O}\left(\min\left\{\left(\frac{M_1}{\mu}\right)^{1/2}, \left(\frac{M_2 R}{\mu}\right)^{2/7}\right\}\left\lceil \log\left(\frac{M_1 R^2}{\max\left\{\mu^3/M_2^2, \varepsilon\right\}}\right)\right\rceil + \log\left\lceil \log\left(\frac{\mu^3/M_2^2}{\varepsilon}\right)\right\rceil\right).$$

Замечание 4. Стоит, однако, отметить, что если вместо сильной выпуклости и достаточной гладкости предполагать *самосогласованность* оптимизируемой функции, то, используя специальную локальную норму Дикина [35]

$$\|u\|_x = \left\langle u, \nabla^2 f(x) u \right\rangle^{1/2},$$

в классе методов 2-го порядка можно улучшить рассматриваемую оценку числа итераций [54, 223]:

$$\mathrm{O}\left(f\left(x^0\right) - f\left(x_*\right) + \log\left\lceil \log\left(1/\varepsilon\right)\right\rceil\right).$$

Введем $g(t) = f(w + t\mathrm{v})$. Самосогласованность $f(x)$ означает, что для любых w и v справедливо неравенство

$$\left|g'''(t)\right| \le 2\left(g''(t)\right)^{3/2}$$

для всех t, причем множитель 2 здесь выбран для определенности. Отметим, что в описанном подходе константы имеют «физическую» размерность, поэтому в отличие от формул, встречавшихся до настоящего момента, в оставшейся части пособия не стоит пытаться проверять корректность формул из соображений размерности [37, 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 также можно достичь следующей оценки числа итераций [54, 223]:

$$\mathrm{O}\left(\sqrt{\nu}\ln\left(\nu/\varepsilon\right)\right).$$

В подавляющем большинстве случаев удается конструктивно показать, явно построив соответствующий самосогласованный барьер, что $\nu \le \mathrm{O}(n)$. ■

◊ В продолжение темы, затронутой в замечании 2, 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, посвященных развитию *методов внутренней точки*/*методов внутренних штрафов* (*барьеров*) [54, 223]. 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 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 [54, 223]. Далее исходная задача заменяется однопараметрическим семейством задач (*метод продолжения по параметру*, см., например, [41, § 5.3])

$$t\langle c, x\rangle + F(x) \to \min_{x},$$

где в конечном итоге нужно, чтобы $t \to \infty$. В упомянутом цикле работ А. С. Немировского и Ю. Е. Нестерова был предложен следующий вариант метода Ньютона:

$$t^{k+1} = \left(1 + \frac{1}{13\sqrt{\nu}}\right) t^k,\ x^{k+1} = x^k - \left[\nabla^2 F\left(x^k\right)\right]^{-1}\left(t^{k+1}\cdot c + \nabla F\left(x^k\right)\right),$$

который после $\mathrm{O}\left(\sqrt{\nu}\ln\left(\varepsilon^{-1}\right)\right)$ 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 $\mathrm{O}\left(\sqrt{\nu}\ln(\nu)\right)$ итераций такого же типа, см., например, [119, п. 5.3].

Заметим, что исходную постановку задачи на самом деле можно завязать на более привычные по данному пособию постановки задачи. В частности:

$$f(x) \to \min_{x}$$

можно эквивалентным образом переписать как

$$y \to \min_{f(x)\le y}.$$

Описанный выше подход лежит в основе пакета выпуклой оптимизации CVX [286], 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* [75, 223, 225], см. также замечание 3.1,

$$f(x) = \min_{y:\, A\begin{bmatrix} x \\ y \end{bmatrix} = b,\, \begin{bmatrix} x \\ y \end{bmatrix} \in K} \langle c, x \rangle + \langle d, y \rangle,$$

где выпуклый конус K – есть прямое произведение конусов трех (канонических) типов: $\mathbb{R}^n_+$, S^n_+, L^n_2, т. е. неотрицательного ортанта, конуса неотрицательно определенных матриц и лоренцовского конуса (ice cream cone). Причем, как было показано А. С. Немировским и А. Бен-Талем в 1998 г. [223, 225], лоренцовский конус L^n_2 можно заменить конусом $\mathbb{R}^{O\left(n \ln\left(\varepsilon^{-1}\right)\right)}_+$. 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 [170, 225].

Заметим, что в начале 90-х годов XX века А. С. Немировский был с длительным визитом у С. Бойда, что, по-видимому, впоследствии и вдохновило С. Бойда с учениками на создание пакета CVX.

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 еще в самых первых работах. В частности, были построены следующие n-самосогласованные барьеры:

$$F_{\mathbb{R}^n_+}(x) = -\sum_{i=1}^{n} x_i \ln x_i,\ F_{S^n_+}(X) = -\ln\det(X).$$

Таким образом и появился один из самых популярных сейчас пакетов CVX решения задач выпуклой оптимизации умеренных размеров – до десятков тысяч переменных [286]. На наш взгляд, CVX имеет наилучшую реализацию для использования вместе с MatLab и не очень удачную реализацию для работы с Python. Отметим также другие пакеты, которые на протяжении многих лет являются помощниками специалистов в области

численных методов оптимизации: CPLEX, MOSEK [289, 290]. И, конечно, нельзя не упомянуть открытую программную библиотеку Google, заточенную под задачи машинного обучения: TensorFlow [292].

Заинтересовавшемуся в замечании 2 и последующем тексте читателю можно рекомендовать ознакомиться с [54, гл. 4], [119, п. 5.3], [223, п. 2.3, 2.5, 3.2] для более глубокого погружения в затронутые здесь темы. Особо отметим то, насколько неожиданными порой бывают выпуклые функции, допускающее явно выписываемое обобщенное представление Фурье–Моцкина [225]. ◊

Также, как и для методов 1-го порядка (см. § 5), для методов 2-го порядка и выше можно рассматривать их универсальные композитные варианты [171, 172], можно рассматривать работу методов в условиях наличия шума и неточностей, возникающих при решении вспомогательных задач на каждой итерации [102, 164], также можно переносить и завязанные на наличие шумов конструкции, например, конструкцию *минибатчинга* [164].

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 эквивалента задаче умножения двух матриц такого же по порядку размера [45, гл. 31], что типично в n раз дороже, чем осуществление шага метода типа градиентного спуска (умножение матрицы на вектор).

◊ На самом деле это не так. Умножение двух матриц $n \times n$ современными алгоритмами может быть осуществлено за время $O\left(n^{2.37}\right)$, см. [64, 192] 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 [164]. Другая идея заключается в отказе от обращения матрицы Гессе на итерации, и вместо этого предлагается использовать информацию о собственном векторе, отвечающем наименьшему собственному значению [82, 120]. Для приближенного вы-

численния такого вектора вполне достаточно уметь умножать матрицу Гессе на произвольный вектор

$$\nabla^2 f(x)\mathrm{v} \approx \frac{\nabla f(x+\tau\mathrm{v}) - \nabla f(x)}{\tau},$$

что может быть сделано с помощью автоматического дифференцирования за то же по порядку время, что и вычисление градиента [106, 246]. Эта идея сейчас активно развивается в связи с поиском наиболее эффективных методов обучения глубоких нейронных сетей [31, 57, 87, 90, 120].

Перспективной также представляется довольно старая идея глобализации сходимости [34, 41, 132, 246]: спуск в область квадратичной сходимости с помощью методов типа градиентного спуска (с дешевыми 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 $\left\|\nabla f\left(x^k\right)\right\|_2 \ll 1$. Если оно выполняется, то делать «пристрелочный» шаг метода Ньютона. Если в результате такого шага выполняется еще и условие

$$\left\|\nabla f\left(x^{k+1}\right)\right\|_2 \ll \left\|\nabla f\left(x^k\right)\right\|_2^{3/2},$$

то продолжать делать шаги метода Ньютона, каждый раз проверяя это условие. Если хотя бы одно из этих условий не выполняется, то вернуться к методу типа градиентного спуска.

Надеемся, что данное пособие вызовет желание поработать в описанных в нем направлениях!

Литература

1. *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.
2. *Аникин А.С., Гасников А.В., Двуреченский П.Е., Тюрин А.И., Чернов А.В.* Двойственные подходы к задачам минимизации сильно выпуклых функционалов простой структуры при аффинных ограничениях // ЖВМ и МФ. – 2017. – Т. 57, № 8. – С. 1270–1284.
3. *Антипин А.С.* Минимизация выпуклых функций на выпуклых множествах с помощью дифференциальных уравнений // Дифференциальные уравнения. – 1994. – Т. 30, № 9. – С. 1395–1375.
4. *Бакушинский А.Б., Кокурин М.Ю.* Итерационные методы решения некорректных операторных уравнений с гладкими операторами. – М.: Едиториал УРСС, 2002. – 192 с.
5. *Баймурзина Д.Р., Гасников А.В., Гасникова Е.В., Двуреченский П.Е., Ершов Е.И., Кубентаева М.Б., Лагуновская А.А.* Универсальный метод поиска равновесий и стохастических равновесий в транспортных сетях // ЖВМ и МФ. – 2019. – Т. 59 (в печати). – URL: https://arxiv.org/ftp/arxiv/papers/1701/1701.02473.pdf
6. *Бадриев И.Б., Задворнов О.А.* Итерационные методы решения вариационных неравенств в гильбертовых пространствах. – Казань: Казанский государственный университет, 2007. – 152 с.
7. *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. – № 9 (в печати). – URL: https://arxiv.org/ftp/arxiv/papers/1701/1701.03821.pdf
8. *Бирюков А.Г.* Методы оптимизации. Условия оптимальности в экстремальных задачах. – М.: МФТИ, 2010. – 225 с.
9. *Бирюков С.И.* Оптимизация. Элементы теории и численные методы. – М: МЗ-Пресс, 2003. – 248 с.
10. *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.
11. *Васильев Ф.П.* Методы оптимизации. Т. 1. – М.: МЦНМО, 2011. – 620 с.
12. *Васильев Ф.П.* Методы оптимизации. Т. 2. – М.: МЦНМО, 2011. – 433 с.
13. *Воронцова Е.А., Гасников А.В., Горбунов Э.А.* Ускоренные спуски по случайному направлению с неевклидовой прокс-структурой // Автоматика и Телемеханика. – 2019 (в печати). – URL: https://arxiv.org/pdf/1710.00162.pdf
14. *Воронцова Е.А., Гасников А.В., Иванова А.С., Нурминский Е.А.* Поиск равновесия по Вальрасу и централизованная распределённая оптимизация с точки зрения современных численных методов выпуклой оптимизации на примере задачи распределения ресурсов // Сиб. ЖВМ. – 2018 (подана).
15. *Гасников А.В.* Эффективные численные методы поиска равновесий в больших транспортных сетях: дис… д.ф.-м.н.: 05.13.18. М.: МФТИ, 2016. – 487 с.

16. *Гасников А.В., Гасникова Е.В., Нестеров Ю.Е., Чернов А.В.* Об эффективных численных методах решения задач энтропийно-линейного программирования // ЖВМ и МФ. – 2016. – Т. 56, № 4. – С. 523–534.
17. *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.
18. *Гасников А.В., Двуреченский П.Е., Нестеров Ю.Е.* Стохастические градиентные методы с неточным оракулом // Труды МФТИ. – 2016. – Т. 8, № 1. – С. 41–91.
19. *Гасников А.В., Двуреченский П.Е., Стонякин Ф.С., Титов А.А.* Адаптивный проксимальный метод для вариационных неравенств // ЖВМ и МФ. – 2019. – Т. 59 (в печати). – URL: https://arxiv.org/pdf/1804.02579.pdf
20. *Гасников А.В., Двуреченский П.Е., Усманова И.Н.* О нетривиальности быстрых (ускоренных) рандомизированных методов // Труды МФТИ. – 2016. – Т. 8, № 2. – С. 67–100.
21. *Гасников А.В., Жуковский М.Е., Ким С.В., Носков Ф.А., Плаунов С.С., Смирнов Д.А.* Вокруг степенного закона распределения компонент вектора PageRank // Сиб. ЖВМ. – 2017–2018. – URL: https://arxiv.org/ftp/arxiv/papers/1701/1701.02595.pdf
22. *Гасников А.В., Камзолов Д.И., Мендель М.А.* Основные конструкции над алгоритмами выпуклой оптимизации и их приложения к получению новых оценок для сильно выпуклых задач // Труды МФТИ. – 2016. – Т. 8, № 3. – С. 25–42.
23. *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.
24. *Гасников А.В., Нестеров Ю.Е.* Универсальный метод для задач стохастической композитной оптимизации // ЖВМ и МФ. – 2018. – Т. 58, № 1. – С. 51–68.
25. *Гилл Ф., Мюррей У., Райт М.* Практическая оптимизация. – М.: Мир, 1985. – 509 с.
26. *Гловински Р., Лионс Ж.Л., Тремольер Р.* Численное исследование вариационных неравенств. – М.: Мир, 1979. – 574 с.
27. *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.
28. *Голуб Дж., Ван Лоун Ч.* Матричные вычисления. – М.: Мир, 1999. – 548 с.
29. *Горнов А.Ю.* Вычислительные технологии решения задач оптимального управления. – Новосибирск: Наука, 2009. – 277 с.
30. *Граничин О.Н., Поляк Б.Т.* Рандомизированные алгоритмы оценивания и оптимизации при почти произвольных помехах. – М.: Наука, 2003. – 291 с.
31. *Гудфеллоу Я., Бенджио И., Курвиль А.* Глубокое обучение. – ДМК Пресс, 2017. – 652 с.
32. *Данскин Дж.М.* Теория максмина. – М.: Советское радио, 1970. – 200 с.
33. *Демьянов В.Ф., Малоземов В.Н.* Введение в минимакс. – М.: Наука, 1972. – 368 с.

34. *Денис Дж., Шнабель Р.* Численные методы безусловной оптимизации и решения нелинейных уравнений. – М.: Мир, 1988. – 440 с.
35. *Дикин И.И.* Метод внутренних точек в линейном и нелинейном программировании. – М.: КРАСАНД, 2010. – 120 с.
36. *Жадан В.Г.* Методы оптимизации. Ч. 1–3. – М.: МФТИ, 2015–2017.
37. *Зорич В.А.* Математический анализ задач естествознания. – М: МЦНМО, 2017. – 160 с.
38. *Евтушенко Ю.Г.* Методы решения экстремальных задач и их применение в системах оптимизации. – М.: Наука, 1982. – 482 с.
39. *Евтушенко Ю.Г.* Оптимизация и быстрое автоматическое дифференцирование. – М.: ВЦ РАН, 2013. – 144 с. – URL: http://www.ccas.ru/personal/evtush/p/198.pdf
40. *Ермольев Ю.М.* Методы стохастического программирования. – М.: Наука, 1976. – 240 с.
41. *Измаилов А.Ф., Солодов М.В.* Численные методы оптимизации. – М.: Физматлит, 2005. – 304 с.
42. *Измаилов А.Ф., Третьяков А.А.* Фактор-анализ нелинейных отображений. – М.: Наука, 1994. – 336 с.
43. *Карманов В.Г.* Математическое программирование. – М.: Наука, 1986. – 288 с.
44. *Ким К., Нестеров Ю., Скоков В., Черкасский Б.* Эффективные алгоритмы для дифференцирования и задачи экстремали // Экономика и математические методы. – 1984. – Т. 20. – С. 309–318.
45. *Кормен Т., Лейзерсон Ч., Ривест Р.* Алгоритмы: построение и анализ. – М.: МЦНМО, 2002. – 960 с.
46. *Корпелевич Г.М.* Экстраградиентный метод для отыскания седловых точек и других задач // Экономика и мат. методы. – 1976. – Т. 12, № 4. – С.747–756.
47. *Магарил-Ильяев Г.Г., Тихомиров В.М.* Выпуклый анализ и его приложения. – М.: УРСС, 2011. – 175 с.
48. *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. – 151 с. URL: http://book.etudes.ru/toc/fast-arithmetic/
49. *Моисеев Н.Н.* Численные методы в теории оптимальных систем. – М.: Наука, 1971. – 424 с.
50. *Моисеев Н.Н., Иванилов Ю.П., Столярова Е.М.* Методы оптимизации. – М.: Наука, 1978. – 351 с.
51. *Немировский А.С., Нестеров Ю.Е.* Оптимальные методы гладкой выпуклой оптимизации // ЖВМ и МФ. – 1985. – Т. 25, № 3. – С. 356–369.
52. *Немировский А.С., Юдин Д.Б.* Сложность задач и эффективность методов оптимизации. – М.: Наука, 1979. – 384 с.
53. *Нестеров Ю.Е.* Алгоритмическая выпуклая оптимизация: дис… д.ф.-м.н.: 01.01.07. М.: МФТИ, 2013. – 367 с.
54. *Нестеров Ю.Е.* Введение в выпуклую оптимизацию. – М.: МЦНМО, 2010. – 262 с.
55. *Нестеров Ю.Е.* Метод минимизации выпуклых функций со скоростью сходимости $O\left(1/k^2\right)$ // ДАН АН СССР. – 1983. – Т.269, № 3. – С. 543–547.
56. *Никайдо Х.* Выпуклые структуры и математическая экономика. – М.: Мир, 1972. – 520 с.

57. *Николенко С., Кадурин А., Архангельская Е.* Глубокое обучение. Погружение в мир нейронных сетей. – СПб.: Питер, 2018. – 480 с.
58. *Новиков П.С.* Элементы математической логики. – М.: Наука, 1973. – 401 с.
59. *Нурминский Е.А.* Численные методы решения детерминированных и стохастических минимаксных задач. – Киев: Наукова думка, 1979. – 160 с.
60. *Обен Ж.-П.* Нелинейный анализ и его экономические приложения. – М.: Мир, 1988. – 264 с.
61. *Поляк Б.Т.* Введение в оптимизацию. – М.: Наука, 1983. – 384 с.
62. *Поляк Б.Т.* Градиентные методы минимизации функционалов, решения уравнений и неравенств: дис… канд. физ.-мат. наук. – М.: МГУ, 1963. – 70 с.
63. *Протасов В.Ю.* К вопросу об алгоритмах приближенного вычисления минимума выпуклой функции по ее значениям // Мат. заметки. – 1996. – Т. 59, № 1. С. 95–102.
64. *Разборов А.А.* Алгебраическая сложность. – М.: МЦНМО, 2016. – 32 с.
65. *Стецюк П.И.* Методы эллипсоидов и *r*-алгоритмы. – Кишинэу: Эврика. 2014. – 488 с.
66. *Стонякин Ф.С.* Адаптивный аналог метода Ю.Е. Нестерова для вариационных неравенств с сильно монотонным полем // Сиб. ЖВМ. – 2018 (в печати). – URL: https://arxiv.org/pdf/1803.04045.pdf
67. *Сухарев А.Г., Тимохов А.В., Федоров В.В.* Курс методов оптимизации. – М.: Физматлит, 2005. – 368 с.
68. *Тер-Крикоров А.М., Шабунин М.И.* Курс математического анализа. – М.: БИНОМ. Лаборатория знаний, 2013. – 672 с.
69. *Тыртышников Е.Е.* Методы численного анализа. – М.: МГУ, 2006. – 281 с.
70. *Тюрин А.И.* Адаптивный быстрый градиентный метод в задачах стохастической оптимизации // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1712.00062.pdf
71. *Тюрин А.И.* Зеркальный вариант метода подобных треугольников для задач условной оптимизации // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1705.09809.pdf
72. *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. – 2019 (в печати). – URL: https://arxiv.org/pdf/1711.02747.pdf
73. *Уайлд Д.Дж.* Методы поиска экстремума. – М.: Наука, 1967. – 268 с.
74. *Хачиян Л.Г.* Избранные труды / сост. С.П. Тарасов. – М.: МЦНМО, 2009. – 520 с.
75. *Циглер Г.М.* Теория многогранников. – М.: МЦНМО, 2014. – 568 с.
76. *Червоненкис А.Я.* Компьютерный анализ данных. – М.: Лекции Школы анализа данных Яндекс, 2009. – 260 с.
77. *Шор Н.З.* Методы минимизации недифференцируемых функции и их приложения. – К.: Наукова думка, 1979. – 199 с.
78. *Эванс Л.К., Гариепи Р.В.* Теория меры и тонкие свойства функций. – Новосибирск: Научная книга, 2002. – 206 с.
79. *Яковлев П.А.* [*и др.*]. Алгоритмы локальной минимизации силового поля для трехмерных макромолекул // ЖВМ и МФ. – 2018 (подана).

80. *Agarwal A., Bartlett P.L., Ravikumar P., Wainwright M.J.* Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization // IEEE Transaction of Information. – 2012. – V. 58, N 5. – P. 3235–3249.
81. *Agarwal A., Dekel O., Xiao L.* Optimal algorithms for online convex optimization with multi-point bandit feedback // COLT. –2010. – P. 28–40.
82. *Agarwal N., Allen-Zhu Z., Bullins B., Hazan E., Ma T.* Finding approximate local minima faster than gradient descent // Proceedings of the Forty-Ninth Annual ACM Symposium on the Theory of Computing, 2017.
83. *Agarwal N., Hazan E.* Lower bounds for higher-order convex optimization // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1710.10329.pdf
84. *Allen-Zhu Z.* Personal web page. – URL: http://people.csail.mit.edu/zeyuan/
85. *Allen-Zhu Z.* Katyusha: the first direct acceleration of stochastic gradient methods // arXiv.org e-Print archive. 2016. – URL: https://arxiv.org/pdf/1603.05953.pdf
86. *Allen-Zhu Z.* Katyusha X: Practical momentum method for stochastic sum-of-nonconvex optimization // arXiv.org e-Print archive. 2018. – URL: https://arxiv.org/pdf/1802.03866.pdf
87. *Allen-Zhu Z.* Natasha 2: Faster non-convex optimization than SGD // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1708.08694.pdf
88. *Allen-Zhu Z., Hazan E.* Optimal black-box reductions between optimization objectives // arXiv.org e-Print archive. 2016. – URL: https://arxiv.org/pdf/1603.05642.pdf
89. *Allen-Zhu Z., Hazan E.* Variance reduction for faster non-convex optimization // arXiv.org e-Print archive. 2016. – URL: https://arxiv.org/pdf/1603.05643.pdf
90. *Allen-Zhu Z., Li Y.* Neon 2: Finding local minima via first-order oracle // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1711.06673.pdf
91. *Allen-Zhu Z., Li Y., Oliveira R., Wigderson A.* Much faster algorithm for matrix scaling // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1704.02315.pdf
92. *Allen-Zhu Z., Orecchia L.* Linear coupling: An ultimate unification of gradient and mirror descent // arXiv.org e-Print archive. 2014. – URL: http://arxiv.org/pdf/1407.1537v4.pdf
93. *Allen-Zhu Z., Qu Z., Richtarik P., Yuan Y.* Even faster accelerated coordinate descent using non-uniform sampling // arXiv.org e-Print archive. 2015. – URL: https://arxiv.org/pdf/1512.09103.pdf
94. *Anandkumar A., Ge R.* Efficient approach for escaping higher order saddle points in non-convex optimization // arXiv.org e-Print archive. 2016. – URL: https://arxiv.org/pdf/1602.05908.pdf
95. *Andrei N.* 40 conjugate gradient algorithms for unconstrained optimization. A survey on their definition // ICI Technical Report N 13/08, March 14, 2008. – URL: https://camo.ici.ro/neculai/p13a08.pdf
96. *Andrychowicz M. et al.* Learning to learn by gradient descent by gradient descent // arXiv.org e-Print archive. 2016. – URL: https://arxiv.org/pdf/1606.04474.pdf
97. *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. – 16 p. – URL: https://arxiv.org/ftp/arxiv/papers/1508/1508.00858.pdf

98. *Antipin A.S.* Gradient approach of computing fixed points of equilibrium problems // Journal of Global Optimization. – 2002. – V. 24(3). – P. 285–309.
99. *Arjevani Y.* Limitation on variance-reduction and acceleration schemes for finite sum optimization // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1706.01686.pdf
100. *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.
101. *Arjevani Y., Shamir O., Shiff R.* Oracle complexity of second-order methods for smooth convex optimization // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1705.07260.pdf
102. *Baes M.* Estimate sequence methods: extensions and approximations // arXiv.org e-Print archive. 2009. – URL: http://www.optimization-online.org/DB_FILE/2009/08/2372.pdf
103. *Bansal N., Gupta A.* Potential-function proofs for first-order methods // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1712.04581.pdf
104. *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, N 2. – P. 330–348.
105. *Bayandina A., Dvurechensky P., Gasnikov A., Stonyakin F., Titov A.* Mirror descent and convex optimization problems with non-smooth inequality constraints. LCCC Focus Period on Large-Scale and Distributed Optimization. Sweden, Lund: Springer, 2018 (accepted). – URL: https://arxiv.org/pdf/1710.06612.pdf
106. *Baydin A.G., Pearlmutter B.A., Radul A.A., Siskand J.M.* Automatic differentiation in machine learning: a survey // arXiv.org e-Print archive. 2015. – URL: https://arxiv.org/pdf/1502.05767.pdf
107. *Beck A.* First-order methods in optimization // MOS-SIAM Series on Optimization. – SIAM, 2017.
108. *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.
109. *Bertsekas D.P.* Constrained optimization and Lagrange multipliers methods. – Athena Scientific, 1996.
110. *Bertsekas D.P.* Convex optimization theory. – Athena Scientific, 2009.
111. *Bertsekas D.P., Nedic A., Ozdaglar A.E.* Convex analysis and optimization. – Belmont, Massachusetts: Athena Scientific, 2003.
112. *Bottou L., Curtis F.E., Nocedal J.* Optimization methods for large-scale machine learning // arXiv.org e-Print archive. 2016. – URL: https://arxiv.org/pdf/1606.04838.pdf
113. *Boyd S.* Personal web-page. – URL: http://stanford.edu/~boyd/
114. *Boyd S., Parikh N.* Proximal algorithms // Foundations and Trends in Optimization. – 2014. – V. 1(3). – P. 123–231.
115. *Boyd S., Parikh N., Chu E., Peleato B., Jonathan E.* Distributed optimization and statistical learning via the alternating direction method of multipliers // Foundations and Trends in Machine Learning. – 2010. – V. 3, N 1. – P. 1–122.
116. *Boyd S., Vandenberghe L.* Convex optimization. – Cambridge University Press, 2004. – URL: https://web.stanford.edu/~boyd/cvxbook/
117. *Boyd S., Vandenberghe L.* Introduction to applied linear algebra – vectors, matrices and least squares. – Camb. Univ. Press, 2018. – URL:

https://web.stanford.edu/~boyd/vmls/
118. *Brent R.P.* Algorithms for minimization without derivatives. – Prentice-Hall, 1973.
119. *Bubeck S.* Convex optimization: algorithms and complexity // Foundations and Trends in Machine Learning. – 2015. – V. 8, N 3–4. – P. 231–357. – URL: https://arxiv.org/pdf/1405.4980.pdf
120. *Carmon Y., Duchi J.C., Hinder O., Sidford A.* Accelerated methods for non-convex optimization // arXiv.org e-Print archive. 2016. – URL: https://arxiv.org/pdf/1611.00756.pdf
121. *Carmon Y., Duchi J.C., Hinder O., Sidford A.* Lower bounds for finding stationary points I // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1710.11606.pdf
122. *Carmon Y., Duchi J.C., Hinder O., Sidford A.* Lower bounds for finding stationary points II: First-order methods // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1711.00841.pdf
123. *Carmon Y., Hinder O., Duchi J.C., Sidford A.* "Convex until proven guilty": Dimension-free accelereation of gradient descent to non-convex functions // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1705.02766.pdf
124. *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.
125. *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, N 5. – P. 32–43.
126. *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.
127. *Chang T.-H., Hong M., Wang X.* Multi-agent distributed optimization via inexact consensus ADMM // IEEE Transactions on Signal Processing. – 2015. – V. 63, N 2. – P. 482–487.
128. *Chen G., Teboulle M.* Convergence analysis of a proximal-like minimization algorithm using Bregman functions // SIAM J. Optim. – 1993. – V. 3. – P. 538–543.
129. *Chen Y., Lan G., Ouyang Y.* Accelereated schemed for class of variational inequalities // Math. Prog., Ser. B. – 2017. – V. 165. – P. 113–149.
130. *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.
131. *Cohen M.B. et al.* Almost-linear-time algorithm for Markov chain and new spectral primitives for directed graphs // arXiv.org e-Print archive. 2016. – URL: https://arxiv.org/pdf/1611.00755.pdf
132. *Conn A.B., Gould N.I.M., Toint P.L.* Trust region methods. – Philadelphia: SIAM, 2000.
133. *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.org e-Print archive. 2015. – URL: http://arxiv.org/pdf/1506.02444v3.pdf
134. *Cuturi M., Peyre G.* A smoothed dual formulation for variational Wasserstein problems // SIAM J. Imaging Science. – 2016. – V. 9. – P. 320–343.

135. *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.
136. *d'Aspremont A.* Smooth minimization with approximate gradient // SIAM Journal on Optimization. – 2008. – V. 19, N 3. – P. 1171–1183.
137. *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.
138. *Devolder O.* Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization: PhD thesis. – CORE UCL, March 2013.
139. *Devolder O., Glineur F., Nesterov Y.* Double smoothing technique for large-scale linearly constrained convex optimization // SIAM J. Optim. – 2012. – V. 22, N 2. – P. 702–727.
140. *Devolder O., Glineur F., Nesterov Yu.* First order methods of smooth convex optimization with inexact oracle // Math. Progr. Ser. A. – 2014. – V. 146(1–2). – P. 37–75.
141. *Devolder O., Glineur F., Nesterov Yu.* First order methods with inexact oracle: the smooth strongly convex case // CORE Discussion Paper 2013/16. – 2013. – 35 p. – URL: https://www.uclouvain.be/cps/ucl/doc/core/documents/coredp2013_16web.pdf
142. *Diakonikolas J., Orecchia L.* The approximate gap technique: A unified approach to optimal first-order methods // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1712.02485.pdf
143. *Drori Y., Taylor A.B.* Efficient first-order methods for convex minimization: a constructive approach // arXiv.org e-Print archive. 2018. – URL: https://arxiv.org/pdf/1803.05676.pdf
144. *Drori Y., Teboulle M.* Performance of first-order methods for smooth convex minimization: a novel approach // Mathematical Programming. – 2014. – V. 145(1–2). – P. 451–482.
145. *Drori Y.* The exact information-based complexity of smooth convex minimization // J. Complexity. – 2017. – V. 39. – P. 1–16.
146. *Drusvyatskiy D.* The proximal point method revisited // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1712.06038.pdf
147. *Du S.S., Hu W.* Linear Convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity // arXiv.org e-Print archive. 2018. – URL: https://arxiv.org/pdf/1802.01504.pdf
148. *Duchi J.C.* Introductory lectures on stochastic optimization // IAS/Park City Mathematics Series. – 2016. URL: http://stanford.edu/~jduchi/PCMIConvex/Duchi16.pdf
149. *Duchi J., Ruan F.* Stochastic methods for composite optimization problems // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1703.08570.pdf
150. *Dvurechensky P.* Gradient method with inexact oracle for composite non-convex optimization // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1703.09180.pdf
151. *Dvurechensky P., Gasnikov A.* Stochastic intermediate gradient method for convex Problems with inexact stochastic oracle // JOTA. – 2016. – V. 171(1). – P. 121–145.

152. *Dvurechensky P., Gasnikov A., Gorbunov E.* An accelerated directional derivative method for smooth stochastic convex optimization // EJOR. – 2018 (submitted). arXiv.org e-Print archive. – 2018. – URL: https://arxiv.org/pdf/1804.02394.pdf
153. *Dvurechensky P., Gasnikov A., Gorbunov E.* An accelerated methods for derivative-free smooth stochastic convex optimization // arXiv.org e-Print archive. 2018. – URL: https://arxiv.org/pdf/1802.09022.pdf
154. *Dvurechensky P., Gasnikov A., Kroshnin A.* Computational optimal transport: complexity by accelerated gradient descent is better than by Sinkhorn's algorithms // ICML. – 2018 (accepted). arXiv.org e-Print archive. 2018. – URL: https://arxiv.org/pdf/1802.04367.pdf
155. *Dvurechensky P., Gasnikov A., Lagunovskaya A.* Parallel algorithms and probability of large deviation for stochastic optimization problems // Siberian Journal of Numerical Mathematics. – V. 21, N 1. – 2018. – P. 47–53.
156. *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.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1707.08486.pdf
157. *Dvurechensky P., Gasnikov A., Kamzolov D.* Universal intermediate gradient method for convex problems with inexact oracle // Optimization Methods and Software. – 2017 (submitted). – URL: https://arxiv.org/pdf/1712.06036.pdf
158. *Dvurechensky P., Nesterov Yu., Spokoiny V.* Primal-dual methods for solving infinite-dimensional games // JOTA. – 2015. – V. 7, N 1. – P. 23–51.
159. *Fercoq O., Qu Z.* Restarting accelerated gradient methods with a rough strong convexity estimate // arXiv.org e-Print archive. 2016. – URL: https://arxiv.org/pdf/1609.07358.pdf
160. *Floudas C.A., Pardalos P.M.* Encyclopedia of optimization. – Kluwer Academic Publishers, 2009.
161. *Gasnikov A.V., Kabanikhin S.I., Mohammed A.A.M., Shishlenin M.A.* Convex optimization in Hilbert space with application to inverse problems // Appl. Numer. Math. – 2017 (submitted). – URL: https://arxiv.org/ftp/arxiv/papers/1703/1703.00267.pdf
162. *Ghadimi S., Feyzmahdavian H.R., Johansson M.* Global convergence of heavy-ball method for convex optimization // arXiv.org e-Print archive. 2014. – URL: https://arxiv.org/pdf/1412.7457.pdf
163. *Ghadimi S., Lan G., Zhang H.* Generalized uniformly optimal methods for nonlinear programming // arXiv.org e-Print archive. 2015. – URL: https://pwp.gatech.edu/guanghui-lan/publications/
164. *Ghadimi S., Liu H., Zhang T.* Second-order methods with cubic regularization under inexact information // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1710.05782.pdf
165. *Gilpin A., Pena J., Sandholm T.* First-order algorithms with $O(1/\varepsilon)$ convergence for ε-equilibrium in two-person zero-sum game // Math. Prog. Ser. A. – 2012. – V. 133. – P. 279–298.
166. *Goodfelow I.J., Vintasls O., Saxe A.M.* Qualitatively characterizing neural network optimization problems // ICLR. – 2015. – URL: https://arxiv.org/pdf/1412.6544.pdf
167. *Golub G.H., Van Loan C.F.* Matrix computations. – Baltimor, MD, USA: John Hopkins University Press, 2012.

168. *Gower R.M., Hanzely F., Richtarik P., Stich S.* Accelerated stochastic matrix inversion: general theory and speeding up BFGS rules for faster second-order optimization // arXiv.org e-Print archive. 2018. – URL: https://arxiv.org/pdf/1802.04079.pdf
169. *Gower R.M., Richtarik P.* Randomized quasi-Newton updates and linearly convergent matrix inversion algorithm // arXiv.org e-Print archive. 2016. – URL: https://arxiv.org/pdf/1602.01768.pdf
170. *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
171. *Grapiglia G.N., Nesterov Yu.* Accelerated regularized Newton method for minimizing composite convex functions // CORE Discussion Papers. 2018/10. – 2018. – 23 p. – URL: https://dial.uclouvain.be/pr/boreal/object/boreal%3A196524/datastream/PDF_01/view
172. *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.
173. *Grimmer B.* Convergence Rates for deterministic and stochastic subgradient methods without Lipschitz continuity // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1712.04104.pdf
174. *Guiges V., Juditsky A., Nemirovski A.* Non-asymptotic confidence bounds for the optimal value of a stochastic program // arXiv.org e-Print archive. 2016 – URL: https://arxiv.org/pdf/1601.07592.pdf
175. *Guminov S., Gasnikov A.* Accelerated methods for α-weakly-quasi-convex optimization problems // Global Optim. – 2017 (submitted). – URL: https://arxiv.org/pdf/1710.00797.pdf
176. *Guminov S., Gasnikov A., Anikin A., Gornov A.* A universal modification of the linear coupling method // Optimization methods and software. – 2017 (submitted). – URL: https://arxiv.org/pdf/1711.01850.pdf
177. *Guzman C., Nemirovski A.* On lower complexity bounds for large-scale smooth convex optimization // Journal of Complexity. – 2015. – V. 31. – P. 1–14.
178. *Hanzely F., Richtarik P.* Fastest rates for stochastic mirror descent methods // arXiv.org e-Print archive. – 2018. – URL: https://arxiv.org/pdf/1803.07374.pdf
179. *Hardt M., Ma T.* Identity matters in Deep Learning // arXiv.org e-Print archive. 2016. – URL: https://arxiv.org/pdf/1611.04231.pdf
180. *Hazan E.* Introduction to online convex optimization // Foundations and Trends® in Optimization. – 2016. – V. 2, N 3–4. – P. 157–325.
181. *Hazan E., Kale S.* Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization // JMLR. – 2014. – V. 15. – P. 2489–2512.
182. *Hu C., Pan W., Kwok J.T.* Accelerated Gradient Methods for Stochastic Optimization and Online Learning // NIPS. – 2009. – URL: https://papers.nips.cc/paper/3817-accelerated-gradient-methods-for-stochastic-optimization-and-online-learning.pdf
183. *Hu B., Lessard L.* Dissipativity theory for Nesterov's accelerated method // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1706.04381.pdf

184. *Jaggi M.* Revisiting Frank–Wolfe: Projection-free sparse convex optimization // Proceedings of the 30th International Conference on Machine Learning. – Atlanta, Georgia, USA, 2013. – 12 p.
185. *Jin C., Netrapalli P., Jordan M.I.* Accelerated gradient descent escapes saddle points faster than gradient descent // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1711.10456.pdf
186. *Juditsky A., Nemirovski A.* First order methods for nonsmooth convex large-scale optimization, I, II. – Optimization for Machine Learning. – MIT Press, 2012.
187. *Juditsky A., Nemirovski A., Tauvel C.* Solving variational inequalities with stochastic Mirror-Prox algorithm // Stochastic Systems. – 2011. – V. 1, N 1. – P. 17–58.
188. *Juditsky A., Nesterov Yu.* Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization // Stoch. System. – 2014. – V. 4, N 1. – P. 44–80.
189. *Kabanikhin S.I.* Inverse and ill-posed problems. – De Gruyter, 2012.
190. *Kakde S.M., Shalev-Shwartz S., Tewari A.* On the duality of strong convexity and strong smoothness: learning applications and matrix regularization // arXiv.org e-Print archive. 2009. – URL: http://ttic.uchicago.edu/~shai/papers/KakadeShalevTewari09.pdf
191. *Karimi H., Nutini J., Schmidt M.* Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition // arXiv.org e-Print archive. 2016. – URL: https://arxiv.org/pdf/1608.04636.pdf
192. *Kelner J.A., Orecchia L., Sidford A., Allen-Zhu Z.* A simple, combinatorial algorithm for solving SDD systems in nearly-linear time // Proceeding STOC'13. – P. 911–920.
193. *Kim D., Fessler J.A.* Adaptive restart of the optimized gradient method for convex optimization // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1703.04641.pdf
194. *Kim D., Fessler J.A.* Generalizing the optimized gradient method for smooth convex minimization // arXiv.org e-Print archive. 2016. – URL: https://arxiv.org/pdf/1607.06764.pdf
195. *Kim D., Fessler J.A.* Optimized first-order methods for smooth convex optimization // Math. Prog. Ser. A. – 2016. – V. 159(1–2). – P. 81–107.
196. *Konecny J., McMahan H.B., Ramage D., Richtarik P.* Federated optimization: distributed machine learning for one-device intelligent // arXiv.org e-Print archive. 2016. – URL: https://arxiv.org/pdf/1610.02527.pdf
197. *Lacost-Julien S., Schmidt M., Bach F.* A simpler approach to obtaining $O(1/t)$ convergence rate for the projected stochastic subgradient method // arXiv.org e-Print archive. 2012. – URL: http://arxiv.org/pdf/1212.2002v2.pdf
198. *Lan G.* Personal web-page. – URL: https://pwp.gatech.edu/guanghui-lan/
199. *Lan G.* Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization // Mathematical Programming. – 2015. – V. 149(1). – P. 1–45.
200. *Lan G.* Gradient sliding for composite optimization // Math. Prog. Ser. A and B. – 2016. – 159, N 1–2. – P. 201–235.
201. *Lan G., Lee S., Zhou Y.* Communication-efficient algorithms for decentralized and stochastic optimization // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1701.03961.pdf

202. *Lan G., Monteiro R.D.C.* Iteration-complexity of first-order augmented Lagrangian methods for convex programming // Math. Prog. Ser. A. – 2016. – V. 155. – P. 511–547.
203. *Lan G., Monteiro R.D.C.* Iteration-complexity of first-order penalty methods for convex programming // Math. Prog., Ser. A. – 2013. – V. 138. – P. 115–139.
204. *Lan G., Yang Y.* Accelerated stochastic algorithms for non-convex finite-sum and multi-block optimization // arXiv.org e-Print archive. 2018. – URL: https://arxiv.org/pdf/1805.05411.pdf
205. *Lan G., Zhou Y.* An optimal randomized incremental gradient method // arXiv.org e-Print archive. 2015. – URL: https://arxiv.org/pdf/1507.02000.pdf
206. *Lan G., Zhou Y.* Randomized gradient extrapolation for distributed and stochastic optimization // arXiv.org e-Print archive. 2017. – URL: https://pwp.gatech.edu/guanghui-lan/publications/
207. *Lee J.D., Panageas I., Piliouras G., Simchowitz M., Jordan M.I., Recht B.* First-order methods almost always avoid saddle point // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1710.07406.pdf
208. *Lee J.D., Simchowitz M., Jordan M.I., Recht B.* Gradient descent only converges to minimizers // 29th Annual Conference on Learning Theory (COLT). – 2016. – P. 1246–1257.
209. *Lee Y.-T., Sidford A., Wong S.C.-W.* A faster cutting plane method and its implications for combinatorial and convex optimization // arXiv.org e-Print archive. 2015. – URL: https://arxiv.org/pdf/1508.04874.pdf
210. *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
211. *Lin H., Mairal J., Harchaoui Z.* Catalyst acceleration for first-order convex optimization: from theory to practice // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1712.05654.pdf
212. *Loizou N., Richtarik P.* Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1712.09677.pdf
213. *Lu H., Freund R.M., Nesterov Yu.* Relatively-smooth convex optimization by first order methods, and applications // arXiv.org e-Print archive. 2016. – URL: https://arxiv.org/pdf/1610.05708.pdf
214. *Malitsky Y., Pock T.* A first-order primal-dual algorithm with linesearch // arXiv.org e-Print archive. 2016. – URL: https://arxiv.org/pdf/1608.08883.pdf
215. *Marial J.* Optimization with first-order surrogate functions // ICML, 2013. – URL: https://arxiv.org/pdf/1305.3120.pdf
216. *McMahan B.H., Moore E., Ramage D., Hampson S., Arcas B.A.* Communication-efficient learning of deep networks from decentralized optimization // arXiv.org e-Print archive. 2016. – URL: https://arxiv.org/pdf/1602.05629.pdf
217. *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.

218. *Mordukhovich B.S.* Variational analysis and generalized differentiation. I basic theory, II applications. – Comprehensive studies in mathematics. – Springer, 2006.
219. *Narkiss G., Zibulevsky M.* Sequential subspace optimization method for large-scale unconstrained problems // Tech. report CCIT N 559, EE Dept. – Technion, 2005. – URL: https://ie.technion.ac.il/~mcib/sesop_report_version301005.pdf
220. *Nedic A., Olshevsky A., Shi W.* Achieving geometric convergence for distributed optimization over time-varying graphs // arXiv.org e-Print archive. 2016. – URL: https://arxiv.org/pdf/1607.03218.pdf
221. *Nedic A., Ozdaglar A.* Cooperative distributed multi-agent optimization. In Convex Optimization in Signal Processing and Communications. – Camb. Univ. Press, 2009. P. 340–386.
222. *Nemirovski A.* Information-based complexity of convex programming. – Technion, Fall Semester 1994/95. – URL: http://www2.isye.gatech.edu/~nemirovs/Lec_EMCO.pdf
223. *Nemirovski A.* Lectures on modern convex optimization analysis, algorithms, and engineering applications. – Philadelphia: SIAM, 2015. – URL: http://www2.isye.gatech.edu/~nemirovs/Lect_ModConvOpt.pdf
224. *Nemirovski A.* Prox-method with rate of convergence $O(1/T)$ 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.
225. *Nemirovski A.* What can be expressed via conic quadratic and semidefinite programming // Presentation. – Technion. – 1998. – URL: https://www2.isye.gatech.edu/~nemirovs/CONRUT.pdf
226. *Nemirovski A., Onn S., Rothblum U.G.* Accuracy certificates for computational problems with convex structure // Math. of Operation Research. – 2010. – V. 35, N 1. – P. 52–78.
227. *Nesterov Yu.* Accelerating the cubic regularization of Newton's method on convex problems // Math. Prog. Ser. A. – 2008. – V. 112. – P. 159–181.
228. *Nesterov Yu.* Efficiency of coordinate descent methods on large scale optimization problems // SIAM Journal on Optimization. – 2012. – V. 22, N 2. – P. 341–362.
229. *Nesterov Yu.* Gradient methods for minimizing composite functions // Math. Prog. – 2013. – V. 140, N 1. – P. 125–161.
230. *Nesterov Yu.* How to make the gradients small // Proc. of OPTIMA 88. – 2012. – P. 10–11.
231. *Nesterov Yu.* Implementable tensor methods in unconstrained convex optimization // CORE Discussion Papers. 2018/5. – 2018. – 22 p. – URL: https://alfresco.uclouvain.be/alfresco/service/guest/streamDownload/workspace/SpacesStore/aabc2323-0bc1-40d4-9653-1c29971e7bd8/coredp2018_05web.pdf?guest=true
232. *Nesterov Yu.* Lexicographical differentiation of nonsmooth functions // Math. Prog. – 2005. – V. 104, N 2–3. – P. 669–700.
233. *Nesterov Yu.* Minimizing functions with bounded variation of subgradients // CORE Discussion Papers. 2005/79. – 2005. – 13 p. – URL: http://webdoc.sub.gwdg.de/ebook/serien/e/CORE/dp2005_79.pdf
234. *Nesterov Yu.* Primal-dual subgradient methods for convex problems // Math. Program. Ser. B. – 2009. – V. 120(1). – P. 261–283.

235. *Nesterov Yu.* Smooth minimization of non-smooth function // Math. Program. Ser. A. – 2005. – V. 103, N 1. – P. 127–152.
236. *Nesterov Yu.* Soft clustering by convex electoral model // CORE Discussion paper. – 2018/01. – 20 p. – URL: https://alfresco.uclouvain.be/alfresco/service/guest/streamDownload/workspace/SpacesStore/ff42ec88-4339-4223-b05d-b768c71ef4e6/coredp2018_01web.pdf?guest=true
237. *Nesterov Yu.* Subgradient methods for huge-scale optimization problems // Math. Program. Ser. A. – 2013. – V. 146, N 1–2. – P. 275–297.
238. *Nesterov Yu.* Universal gradient methods for convex optimization problems // Math. Program. Ser. A. – 2015. – V. 152. – P. 381–404.
239. *Nesterov Yu., Polyak B.* Cubic regularization of Newton method and its global performance // Math. Program. Ser. A. – 2006. – V. 108. – P. 177–205.
240. *Nesterov Yu., Shikhman V.* Distributed price adjustment based on convex analysis // Journal Optimization Theory and Applications. – 2017. – V. 172(2). – P. 594–622.
241. *Nesterov Yu., Shikhman* V. Dual subgradient method with averaging for optimal resource allocation // CORE Discussion paper. – 2017/13. – 19 p. – URL: https://dial.uclouvain.be/pr/boreal/object/boreal%3A184242/datastream/PDF_01/view
242. *Nesterov Yu., Shpirko S.* Primal-dual subgradient method for huge-scale linear conic problem // SIAM Journal on Optimization. – 2014. – V. 24, N 3. – P. 1444–1457.
243. *Nesterov Yu., Spokoiny V.* Random gradient-free minimization of convex functions // Foundations of Computational Mathematics. – 2017. – V. 17(2). – P. 527–566.
244. *Nesterov Yu., Stich S.* Efficiency of accelerated coordinate descent method on structured optimization problems // SIAM J. Optim. – 2017. – V. 27, N 1. – P. 110–123.
245. *Niu F., Rechy B., Re C., Wright S.J.* HOGWILD! A lock-free approach to parallelizing stochastic gradient // arXiv.org e-Print archive. 2011. – URL: https://arxiv.org/pdf/1106.5730.pdf
246. *Nocedal J., Wright S.* Numerical optimization. – Springer, 2006.
247. *Ochs P., Fadili J., Brox T.* Non-smooth non-convex Bregman minimization: unification and new algorithms // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1707.02278.pdf
248. *O'Donoghue B., Candes E.* Adaptive restart for accelerated gradient schemes // Foundations of Computational Mathematics. – 2015. – V. 15. – P. 715–732.
249. *Ouyang Y., Chen Y., Lan G., Pasiliao Jr. E.* An accelerated linearized direction method of multipliers // SIAM J. Imaging Science. – 2015. – V. 8, N 1. – P. 644–681.
250. *Peypouquet J.* Convex optimization in normed spaces: theory, methods and examples // arXiv.org e-Print archive. 2014. – URL: http://jpeypou.mat.utfsm.cl/wp-content/uploads/2014/05/Convex-Optimization.pdf
251. *Peyre G., Cuturi M.* Computational optimal transport // arXiv.org e-Print archive. 2018. – URL: https://arxiv.org/pdf/1803.00567.pdf
252. *Polyak B.* History of mathematical programming in the USSR: analyzing the phenomenon // Math. Progr. ser. B. – 2002. – V. 91, N 3. – P. 401–416. – URL: lab7.ipu.ru/files/polyak/Pol-rus-Baikal%2708.pdf
253. *Polyak B., Tremba A.* Solving underdetermined nonlinear equations by Newton-like method // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1703.07810.pdf

254. *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
255. *Richtarik P.* Personal web page. – URL: http://www.maths.ed.ac.uk/~prichtar/
256. *Richtarik P., Takac M.* Stochastic reformulation of linear systems: algorithms and convergence theory // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1706.01108.pdf
257. *Renegar J., Grimmer B.* A simple nearly-optimal restart scheme for speeding-up first order methods // arXiv.org e-Print archive. 2018. – URL: https://arxiv.org/pdf/1803.00151.pdf
258. *Rogozin A., Uribe C., Gasnikov A., Malkovsky N., Nedic A.* Optimal distributed convex optimization on slowly time-varying graphs // arXiv.org e-Print archive. 2018. – URL: https://arxiv.org/pdf/1805.06045.pdf
259. *Roulet V., d'Aspremont A.* Sharpness, restart and acceleration // NIPS, 2017. – P. 1119–1129. – URL: https://arxiv.org/pdf/1702.03828.pdf
260. *Ruder A.* An overview of gradient descent optimization algorithms // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1609.04747.pdf
261. *Scaman K., Bach F., Bubeck S., Lee Y.T., Massoulie L.* Optimal algorithms for smooth and strongly convex distributed optimization in networks // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1702.08704.pdf
262. *Scieur D., d'Aspremont A., Bach F.* Regularized nonlinear acceleration // arXiv.org e-Print archive. 2016. – URL: https://arxiv.org/pdf/1606.04133.pdf
263. *Scieur D., Roulet V., Bach F., d'Aspremont A.* Integration methods and accelerated optimization algorithms // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1702.06751.pdf
264. *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, N 1. – P. 49–54.
265. *Shalev-Shwartz S., Ben-David S.* Understanding Machine Learning: From theory to algorithms. – Cambridge University Press, 2014.
266. *Shalev-Shwartz S., Zhang T.* Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization // ICML. – 2014. – P. 64–72.
267. *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.
268. *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.
269. *Spall J.C.* Introduction to stochastic search and optimization: estimation, simulation and control. – Wiley, 2003.
270. *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.
271. *Sridharan K.* Learning from an optimization viewpoint. – PhD Thesis, Toyota Technological Institute at Chicago, 2011. – URL: http://ttic.uchicago.edu/~karthik/thesis.pdf

272. *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.
273. *Sun H., Hong M.* Distributed non-convex first-order optimization and information processing: Lower complexity bounds and rate optimal algorithms // arXiv.org e-Print archive. 2018. – URL: https://arxiv.org/pdf/1804.02729.pdf
274. *Sutskever I., Martens J., Dahl G., Hinton G.* On the importance of initialization and momentum in deep learning // ICML. – 2013. – P. 1139–1147.
275. *Taylor A.B., Hendrickx J.M., Glineur F.* Smooth strongly convex interpolation and exact worst-case performance of first-order methods // Mathematical Programming. – 2017. – V. 161, N 1–2. – P. 307–345.
276. *Taylor A., Van Scoy B., Lessard L.* Lyapunov functions for first-order methods: tight automated convergence guarantees // arXiv.org e-Print archive. 2018. – URL: https://arxiv.org/pdf/1803.06073.pdf
277. *Tran-Dinh Q., Fercoq O., Cevher V.* A smoothing primal-dual optimization framework for nonsmooth composite convex optimization // SIAM J. Optim. – 2018. – URL: https://arxiv.org/pdf/1507.06243.pdf
278. *Trefethen L.N., Bau D. III* Numerical linear algebra. – SIAM, 1997.
279. *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
280. *Uribe C.A., Dvinskikh D., Dvurechensky P., Gasnikov A., Nedic A.* Distributed computation of Wasserstein barycenters over networks // arXiv.org e-Print archive. 2018. – URL: https://arxiv.org/pdf/1803.02933.pdf
281. *Uribe C.A., Lee S., Gasnikov A., Nedic A.* Optimal algorithms for distributed optimization // arXiv.org e-Print archive. 2017. – URL: https://arxiv.org/pdf/1712.00232.pdf
282. *Wilson A.C., Recht B., Jordan M.I.* A Lyapunov analysis of momentum methods in optimization // arXiv.org e-Print archive. 2016. – URL: https://arxiv.org/pdf/1611.02635.pdf
283. *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
284. *Xu Y.* Accelerated first-order primal-dual proximal methods for linearly constrained composite convex programming // arXiv.org e-Print archive. 2016. – URL: https://arxiv.org/pdf/1606.09155.pdf
285. *Yurtsever A., Tran-Dinh Q., Cevher V.* A universal primal-dual convex optimization framework // NIPS. – 2015. – P. 3150–3158. – URL: https://arxiv.org/pdf/1502.03123.pdf
286. http://cvxr.com/cvx/
287. https://github.com/amkatrutsa/MIPT-Opt
288. https://habrahabr.ru/company/intel/blog/80342/
289. https://www.ibm.com/analytics/data-science/prescriptive-analytics/cplex-optimizer
290. https://www.mosek.com/
291. http://nbviewer.jupyter.org/github/merkulovdaniil/mipt_optimization/tree/master/
292. https://www.tensorflow.org/

Учебное издание

Гасников Александр Владимирович

СОВРЕМЕННЫЕ ЧИСЛЕННЫЕ МЕТОДЫ ОПТИМИЗАЦИИ.

МЕТОД УНИВЕРСАЛЬНОГО ГРАДИЕНТНОГО СПУСКА

Редакторы: *В. А. Дружинина, И. А. Волкова*, *О. П. Котова*.
Корректор *Н. Е. Кобзева*
Компьютерная верстка: *Н. Е. Кобзева*, *Е. А. Казённова*
Дизайн обложки *Е. А. Казённова*

Подписано в печать 24.04.2018. Формат 60 × 84 $^{1}/_{16}$. Усл. печ. л. 10,4.
Уч.-изд. л. 9,5. Тираж 150 экз. Заказ № 127.

Федеральное государственное автономное образовательное
учреждение высшего образования
«Московский физико-технический институт (государственный университет)»
141700, Московская обл., г. Долгопрудный, Институтский пер., 9
Тел. (495) 408-58-22, e-mail: rio@mipt.ru

Отдел оперативной полиграфии «Физтех-полиграф»
141700, Московская обл., г. Долгопрудный, Институтский пер., 9
Тел. (495) 408-84-30, e-mail: polygraph@mipt.ru